Edward Layer
Krzysztof Tomczyk
Editors

Measurements, Modelling and Simulation of Dynamic Systems

Springer

Measurements, Modelling and Simulation of Dynamic Systems

Edward Layer and Krzysztof Tomczyk (Eds.)

Measurements, Modelling and Simulation of Dynamic Systems

Springer

Prof. Edward Layer
Cracow University of Technology
Faculty of Electrical and Computer Engineering
31-155 Cracow
Warszawska 24
Poland
E-mail: elay@pk.edu.pl

Dr. Krzysztof Tomczyk
Cracow University of Technology
Faculty of Electrical and Computer Engineering
31-155 Cracow
Warszawska 24
Poland
E-mail: ktomczyk@pk.edu.pl

ISBN 978-3-642-04587-5 e-ISBN 978-3-642-04588-2

DOI 10.1007/978-3-642-04588-2

Library of Congress Control Number: 2009937027

Typesetting & Cover Design: Scientific Publishing Services Pvt. Ltd., Chennai, India

Printed in acid-free paper

9 8 7 6 5 4 3 2 1

springer.com

Preface

The development and use of models of various objects is becoming a more common practice in recent days. This is due to the ease with which models can be developed and examined through the use of computers and appropriate software. Of those two, the former - high-speed computers - are easily accessible nowadays, and the latter - existing programs - are being updated almost continuously, and at the same time new powerful software is being developed.

Usually a model represents correlations between some processes and their interactions, with better or worse quality of representation. It details and characterizes a part of the real world taking into account a structure of phenomena, as well as quantitative and qualitative relations. There are a great variety of models. Modelling is carried out in many diverse fields. All types of natural phenomena in the area of biology, ecology and medicine are possible subjects for modelling. Models stand for and represent technical objects in physics, chemistry, engineering, social events and behaviours in sociology, financial matters, investments and stock markets in economy, strategy and tactics, defence, security and safety in military fields. There is one common point for all models. We expect them to fulfil the validity of prediction. It means that through the analysis of models it is possible to predict phenomena, which may occur in a fragment of the real world represented by a given model. We also expect to be able to predict future reactions to signals from the outside world.

There are many ways of the describing a system or its events, which means many ways of constructing a model. We may use words, drawings, graphs, charts, tables, physical models, computer programs, equations and mathematical formulae. In other words, for modelling we can use various methods applying them individually or in parallel. If models are developed by the use of words and descriptions, then the link between cause-and-effect is usually of qualitative character only. Such models are not fully satisfying as the quantitative part of the analysis is missing. A necessary supplement of modelling is the identification of parameters and methods of their measurement. A comprehensive model that includes all these parameters in a numerical form will help us explain the reactions and the behaviours of the objects that are of interest. The model must also enable us to predict the progression of events in the future. Obviously, all those features are linked directly to the accuracy of the model, which in turn depends on the construction of the model and its verifications.

The most common and basic approach to modelling is the identification approach. When using it, we observe actual inputs and outputs and try to fit a model to the observations. In other words, models and their parameters are identified through experiments.

Two methods of identification can be distinguished, namely the active and passive, the latter usually less accurate

The identification experiment lasts a certain period of time. The object under test is excited by the input signal, usually a standard one, and the output is observed. Then we try to fit a model to the observations. That is followed by an estimation of parameters. At this point model quality is verified, and checked whether it satisfies a requirement. If not, we repeat the process taking a more complex model structure into consideration and adjusting its parameters. The model's quality is verified again and again until the result is satisfactory.

In such modelling, difficulties can be expected in two areas and can be related to model structure and parameter estimation. One potential problem is non-linearity of elements or environment during dynamic operation. This can increase the number of difficulties in the development of a model's structure. An estimation of parameters can also be difficult, usually burdened with errors related to interference and random noise in the experiment.

In this book, for modelling we will be using mathematics, especially equations, leading to mathematical models. We will concentrate on models of objects applied and utilized in technology. The described reality and phenomena occurring in it are of analogue character. Their mathematical representation is usually given by a set of equations containing variables, their derivatives and integrals. Having a set with one variable and differentiating it, we can eliminate integrals. The result of this operation is a set of differential equations having one independent variable. Very often time is that independent variable. Such being the case, it is quite convenient to express equations as state equations or transfer functions. Both methods are quite common particularly in the area of technology.

Most commonly, models are sets of linear equations. Their linearity is based on the assumption that either they represent linear objects or that nonlinearities are so small that they can be neglected and the object can be described by linear equations. Such an approach is good enough and well based in many practical cases, and the resulting model accuracy confirmed by verification is satisfactory. Usually verification is carried out for a certain operation mode of a system described by the model. If this mode changes dynamically and is not fixed precisely, model verification may be difficult. In this case verification of the model can be related to signals generating maximum errors. The sense of it is such that the error produced by the application of those signals will always be greater, or at most equal, to the error generated by any other signal. At this point the question must be answered whether signals maximizing chosen error criteria exist and are available during the specific period of time. In such cases, the accuracy of the model should be presented by the following characteristic - maximum error vs. time of input signal duration.

Approximation methods are another popular way of mathematical representation. In this case a model is shown in the form of algebraic polynomials, often orthogonal. These can be transformed into state equations or transfer functions.

The construction of a model is based on experimental data. To obtain data, measurements are carried out, usually supported by personal computer based data acquisition systems or computer aided measuring systems. Dedicated software controls these. Appropriate programs process acquired data. A data acquisition card, which is a part of the system, must be plugged-in into a USB or PCI slots. A computer structure, its elements and operation are presented in Chapters 1 and 2. Quite often signals measured are distorted by noise. Problems related to noise reduction are discussed in Chapter 3. In Chapter 4, a number of mathematical methods for modelling are presented and discussed. The application of the powerful graphical programming LabVIEW software for models development and analysis is also included in the chapter. Finally, in the same Chapter 4 the use of the MATLAB package for the black-box type and Monte-Carlo method of identification is discussed. The last chapter covers the problems of model accuracy for some difficult cases, when input signals are dynamically varying and are of undetermined and unpredictable shapes. A solution to these problems is based on the maximum errors theory. Particularly, this theory creates a possibility for elaborating and establishing the calibration methods and hierarchies of accuracy for dynamic measuring systems, which have not been worked out so far. For a detailed consideration the examples of the integral-squared error and the absolute value of error are discussed and explained in details.

This book is directed towards students as well as industrial engineers and scientists of many engineering disciplines who use measurements, mathematical modelling techniques and computer simulation in their research. The authors hope that this book may be an inspiration for further projects related to modelling and model verification and application.

Acknowledgments

We wish to acknowledge a continuous support and encouragement of the Vice-Chancellor of the Cracow University of Technology Professor Kazimierz Furtak and the Head of Faculty of Electrical and Computer Engineering Professor Piotr Drozdowski.

We would also like to express special thanks to Ilona, Beata, Magdalena, and Piotr - our wives and children, for her patience and support during the writing of this book.

Edward Layer
Krzysztof Tomczyk

Contents

Chapter 1
Introduction to Measuring Systems

A development of mathematical models is based, among others, on some data. These can be more or less reliable. In general, their verification can only be carried out when the model quality is checked up. If it comes to the point that the quality of the model developed does not satisfy the requirements, i.e. that there is a significant difference between the model and the part of reality represented by it, then the data applied for modeling are incorrect or incomplete. In practice, data for the model development originate from measurements of some signals involved. Such measurements are usually carried out with the use of special measuring systems. In general, such systems can be quite similar to each other or have some differences at some points; all depend on application. However, it can be noticed that modern systems have many common elements and components. Starting from the input signal element i.e. a sensor first of all, we can further list components of digital processing and signal conditioning, recording components, output elements and storage devices. These elements and components can be identified in measuring systems that process and measure very different signals of various amplitudes, dynamic properties, forms of energy transferred or various transmitted frequency bands.

A measured quantity which is acting on the sensor, is of the analogue form like all other phenomena in the real world surrounding us. Since computer-aided measuring systems operate using discrete signals only, hence analogue input signals to these systems must be in the first place converted into discrete signals.

A basic measuring system is shown in Fig. 1.1. It includes the conversion of an analogue signal into the digital one, mathematical processing of the signal and its recording. These basic blocks of operation can be seen in all types of measuring systems. Hence, their construction, principle of operation, purpose and application are discussed further in the text. Having been acquainted with them, the reader can make a correct synthesis of basically any measuring system, which carries out various measurements, data collection and recording, also for other aims than the modelling and model development.

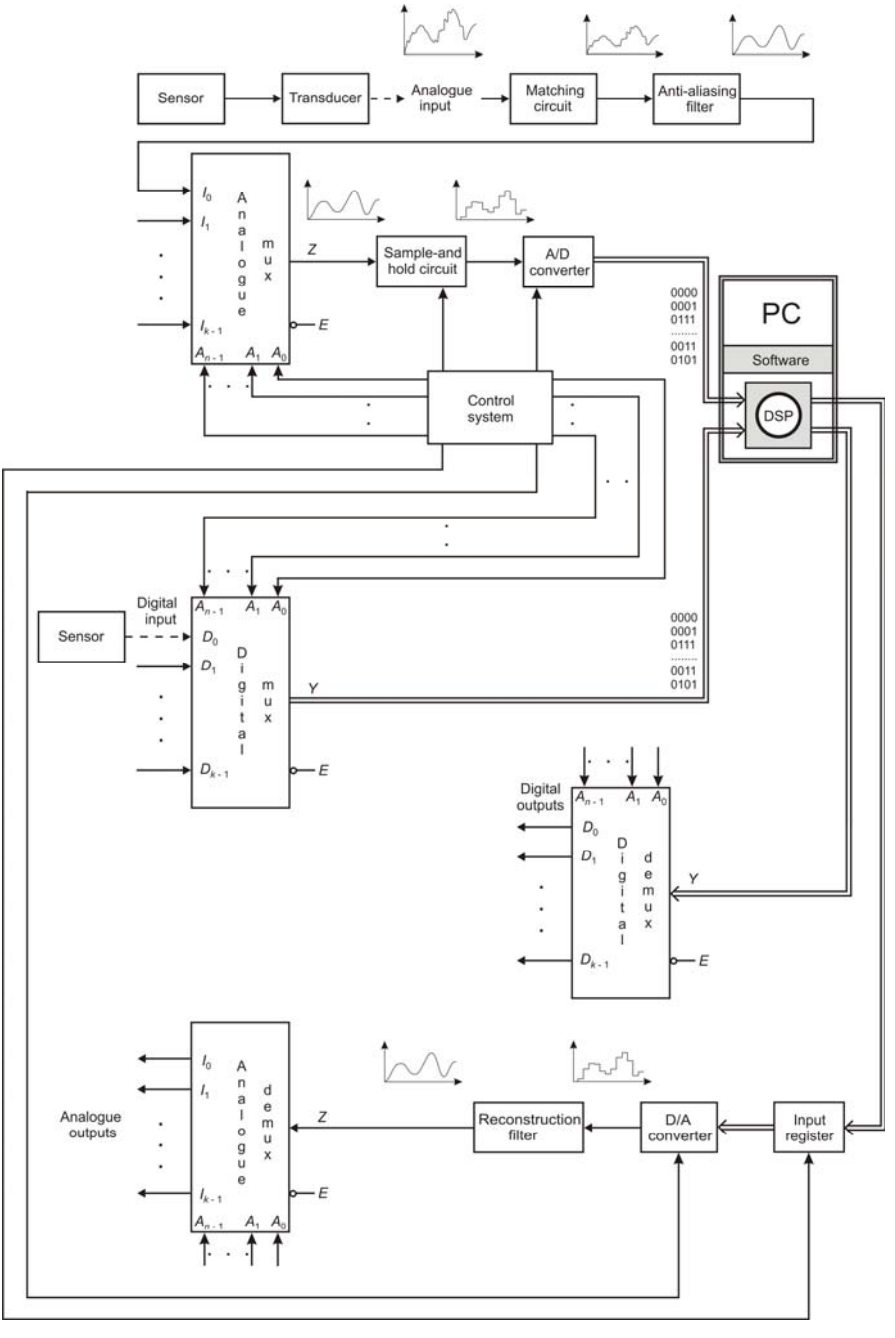

Fig. 1.1 Block diagram of a basic measuring system

1.1 Sensor

A sensor is a device that detects a change in an analogue quantity, which is to be measured, and turns into another physical quantity. This one in turns is converted usually into current or voltage by a transducer.

We can distinguish parametric and self-generating sensors. In case of the former, change of the measured quantity is followed by a change of a parameter of electric circuit, for example resistance, capacitance, self-inductance or mutual inductance. In case of self-generating sensors, a measured quantity is usually changed directly into voltage, current or electric charge.

There are also coding sensors. Their digitized output goes directly towards the digital channel of a measuring system.

1.2 Transducer

The type of a transducer depends on the kind of the output signal from a sensor. Most often bridge circuits or half-bridge circuits are applied for this purpose. They operate in connection with parametric sensors, for example strain gauges that are used for the measurement of dynamically changing strain. Other types of transducer measuring circuits are applied in connection with capacitive and inductive transducers that are used for a measurement of pressure difference and linear displacement.

1.3 Matching Circuit

A matching circuit is applied for adjusting the range of measuring channel and its input impedance. Its key element is the amplifier. A very high value of the amplifier input impedance protects the sensor from loading. The adjustable gain makes possible to select a required range appropriate for a measured signal. There are three most important groups of these amplifiers:

- non-programmable amplifiers with the gain adjustable by a change of feedback loop parameters
- programmable amplifiers with the digitally programmable gain controlled and adjustable by a control system
- amplifiers with optocouplers, having isolated circuit's output from its input.

1.4 Anti-aliasing Filter

The forth block of the measuring system (Fig. 1.1) is a low-pass anti-aliasing filter. It removes all harmonics of the measured signal that exceed the Nyquist frequency. An aliasing error is produced when the sampling frequency is not at least twice as high as the highest measured signal frequency and the overlap

between those frequencies appears. Fig. 1.2 shows spectrum diagrams explaining the overlap and the causes of aliasing error.

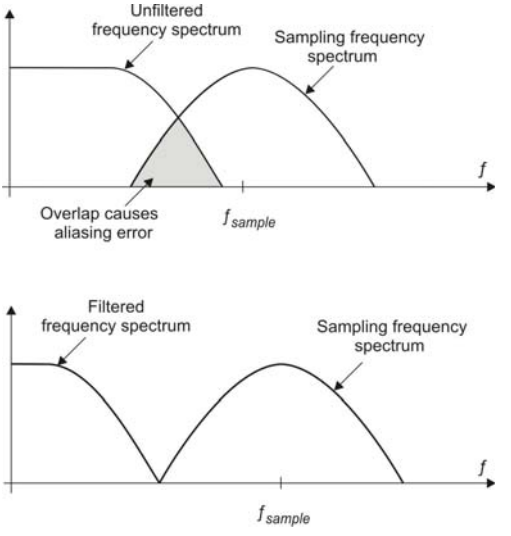

Fig. 1.2 Illustration of the aliasing phenomenon

The anti-aliasing filter should have the following properties: flat passband, sharp cut-off characteristic and low distortion in the passband.

Filter are characterized by their frequency response characteristic $K(\omega)$. Butterworth and Tchebychev low-pass filters are commonly used as anti-aliasing filter. Frequency response characteristic $K(\omega)$ of Butterworth low-pass filter is given by

$$\left|K(\omega)\right| = \frac{k}{\sqrt{1+\left(\frac{\omega}{\omega_c}\right)^{2n}}} \tag{1.1}$$

while for Tchebychev filter we have

$$\left|K(\omega)\right| = \frac{k}{\prod_n \sqrt{\left[1-b_n\left(\frac{\omega}{\omega_c}\right)^2\right]^2 + a_n^2\left(\frac{\omega}{\omega_c}\right)^2}} \tag{1.2}$$

where n, k, ω_c are order, gain and cut-off frequency of the filter, a_n, $b_n \in \Re$.

According to Shannon theorem, the relation between highest frequencies in the analogue signal being sampled and the sampling frequency should be as follows

$$f_{sample} \geq 2 f_{max(\text{analogue})} \tag{1.3}$$

If the analogue signal containing a spectrum of different frequency components is sampled, we obtain a series of impulses modulated in amplitude. In the frequency domain it corresponds to a spectrum of harmonic, as shown in Fig. 1.3.

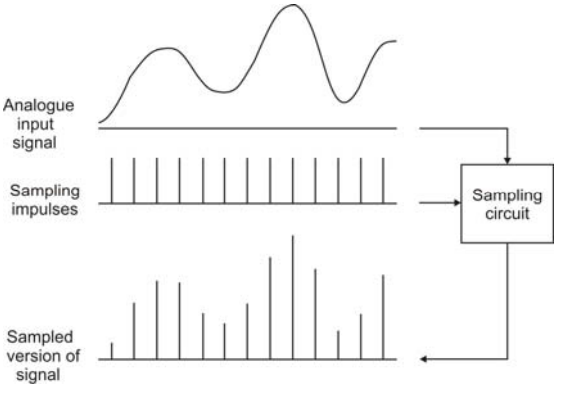

Fig. 1.3 Illustration of the sampling process

If the maximum frequency $f_{max(\text{analogue})}$ of a signal increases, the individual spectra will widen and begin to overlap. In effect, the original signal cannot be accurately reproduced. Thus, for accurate reproduction of a signal containing the frequency up to $f_{max(\text{analogue})}$ the sampling rate must be grater than, or equal to $2 f_{max(\text{analogue})}$. The condition can be achieved by sufficiently increasing the sampling frequency. Unfortunately, the maximum of this frequency is usually limited by the performance of the A/D converter, which is the block next in line in the measuring system discussed. If sampling frequency cannot be adequately increased, a low-pass anti-aliasing filter must be used in order to truncate the signal spectrum to the desired value of $f_{max(\text{analogue})}$ for a given sample frequency.

A conversion of a continuous analogue signal consists of three steps, namely: sampling, quantization (to digitize the value of a signal) and coding of the resulting signal.

Sampling is a digital process carried on in time and related to the argument of the input signal. Samples of the input signal values are collected in clearly defined intervals.

Quantization is a process of assigning a finite number of magnitude levels to each sample of the converted signal. Each magnitude is denoted by some digital numbers, from zero to the maximum value of conversion range.

Coding simply means the representation of quantized value of the signal by a selected code, most often a natural binary one or Gray code.

1.5 Multiplexers/Demultiplexers

A digital multiplexer is a multi-input and single-output switch, which selects one of many data inputs $D_0, D_1, ..., D_{k-1}$, and sends it to the single output Y.

A multiplexer has k data inputs, n address inputs $A_0, A_1, ..., A_{n-1}$, usually $k = 2^n$, one output and one enable (strobe) control input $\overline{E}$. Fig. 1.4 shows the $8-$input digital multiplexer as an example.

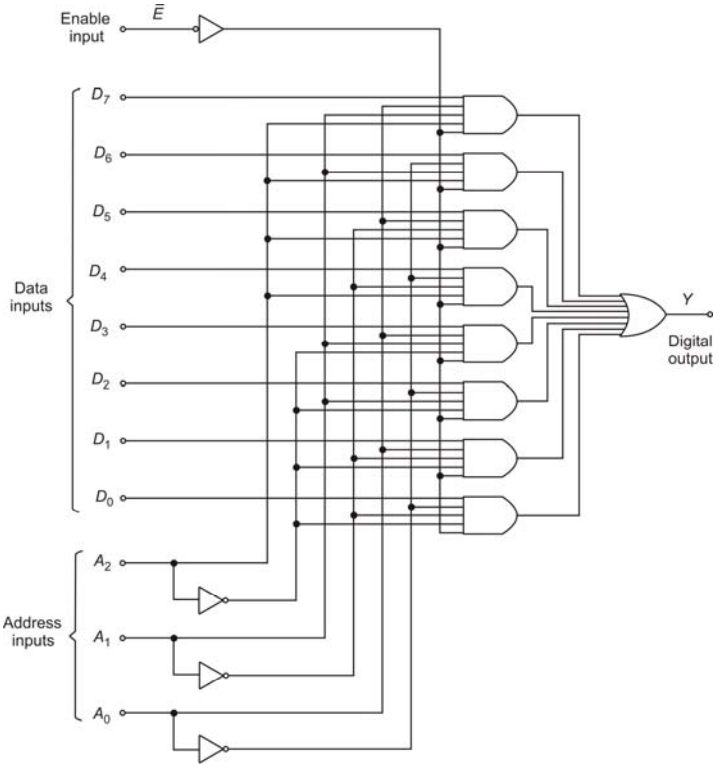

Fig. 1.4 Logic circuit of digital multiplexer

If there is the zero logic signal at the enable input, then the output *Y* receives particular logic state (usually zero), which is independent from the input states *D* and *A*. The binary coded address determines which input signal will appear at the output line. This signal is divided into segments, which are kept at the output as long as the input address will not be changed. However, the input addresses change in a rotating, repeating sequence most often. Multiplexers have usually 4, 8, or 16 parallel inputs. The logic circuit of the demultiplexer is shown in Fig. 1.5

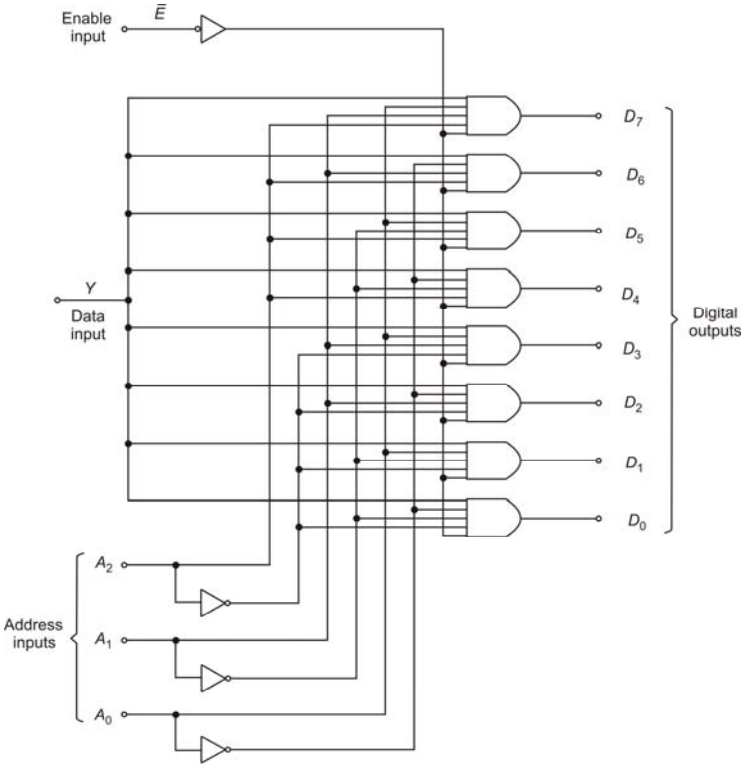

Fig. 1.5 Logic circuit of digital demultiplexer

Analogue multiplexers/demultiplexers – Fig. 1.6 are made by replacing the AND gates through the digitally controlled analogue gates made e.g. in CMOS technology, as shown in Fig. 1.7. Analogue switches are based on field-effect transistors with CMOS insulated gates. They make possible the bidirectional operation and they switch analogue voltages of the peak-to-peak value up to 15V. Analogue switches have a small resistance, if control input is high, and a very high resistance, if this input is low.

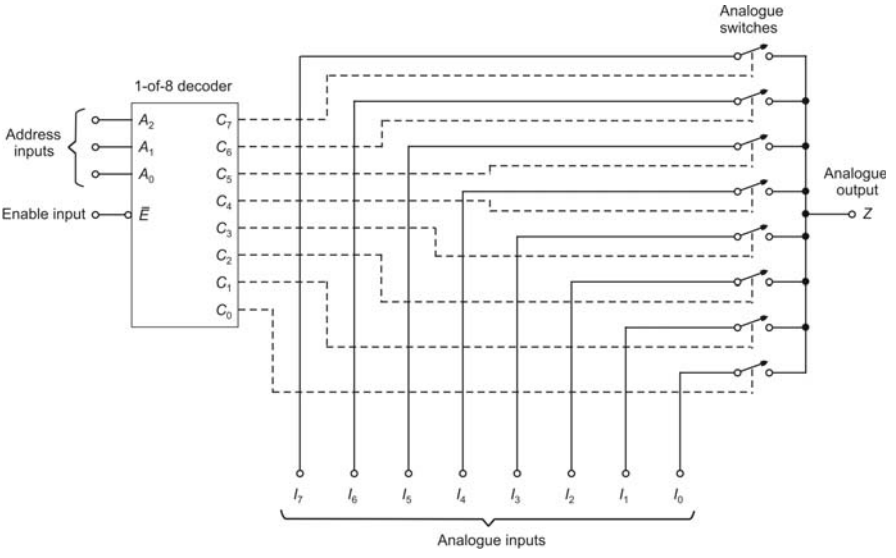

Fig. 1.6 Logic circuit of analogue multiplexer/demultiplexer

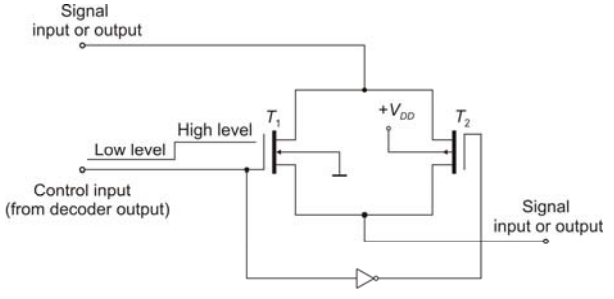

Fig. 1.7 CMOS analogue switch

1.6 Sample-and-Hold Circuit

A sample-and-hold circuit (S/H) samples and temporarily stores the value of an analogue signal for subsequent processing. After filtering and sampling, the sampled level of the signal must be frozen until the A/D converter digitizes it and the next sampling occurs. For this reason, the S/H circuit is switched on for a short period, first into the sample mode and then into the hold mode. This switching is controlled by the voltage control $V_{control}$ in a following way

$$\begin{aligned} V_{out}(t) &= V_{in}(t) \quad \text{if} \quad V_{control} = 0 \Leftrightarrow \text{sample} \\ V_{out}(t) &= V_{in}(t_{0/1}) \quad \text{if} \quad V_{control} = 1 \Leftrightarrow \text{hold} \end{aligned} \tag{1.4}$$

The sample-and-hold operation results in a stairs waveform that approximates the analogue signal. Fig. 1.8 shows a sample-and-hold operation while its circuit is shown in Fig. 1.9.

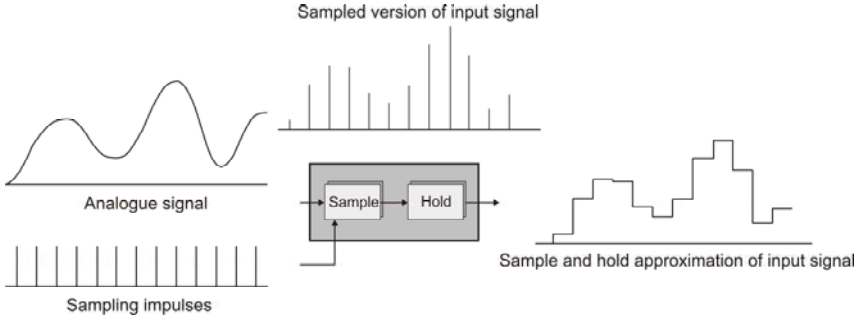

Fig. 1.8 Sample-and-hold operation

When $V_{control} = 0$ the capacitor is charged and the interval is known as the acquisition period or 'aperture time'. Its value is of the order of $0.5 - 20\,\mu s$ varying for different types of the S/H circuit. It depends upon the magnitude of the input voltage. When $V_{control}$ is switched to 1, the S/H circuit is put on hold and the output signal $V_{out}(t)$ equals to the input signal $V_{in}(t_{0/1})$.

The output voltage is digitized by the A/D converter. When digitized, the charge of the hold-capacitor begins to decay causing the drift in the S/H's output voltage. The use of the large hold-capacitors will minimize the output voltage drift and extend the acquisition time. In practice, inclusion of high resistance input amplifiers, which reduce discharge of the capacitors, can make an improvement.

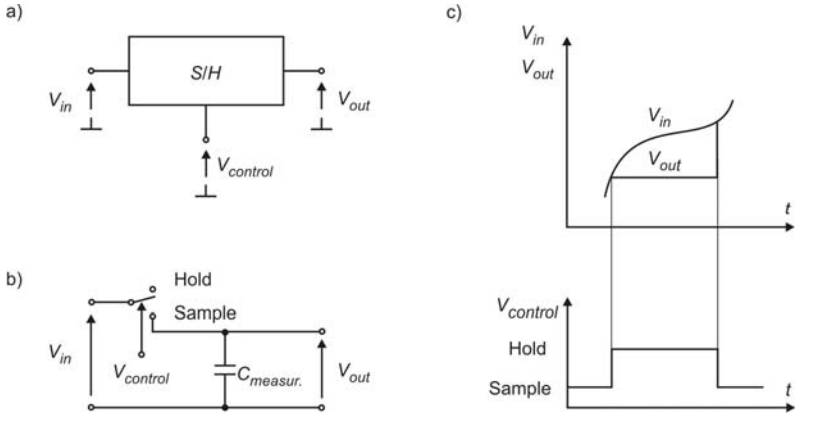

Fig. 1.9 Sample-and-hold circuit a) Terminals notation b) Principle of operation c) Exemplary signals

1.7 Analogue-to-Digital Conversion

Analogue-to-digital conversion is the process of converting the analogue voltage output to a series of binary codes that represent the magnitude of this voltage at each of the sample times. The principle of operation of A/D conversion is shown in Fig. 1.10.

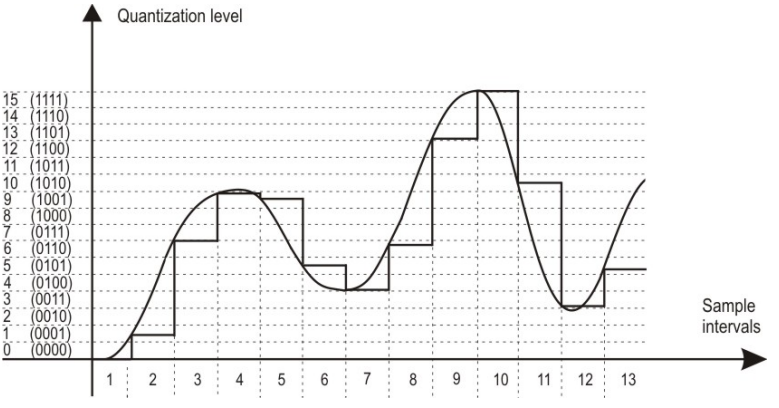

Fig. 1.10 A/D conversion

1.7.1 A/D Converter with Parallel Comparison

Converters of this type are based on the direct comparison of the analogue voltage with one of 2^n reference sectors. Fig. 1.11 presents the logic circuit of an example of the $n-$bit A/D converter with parallel comparison.

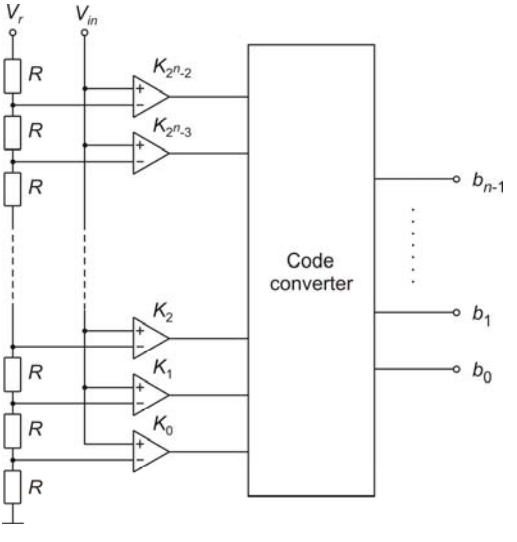

Fig. 1.11 Logic circuit of $n-$bit A/D converter with direct parallel comparison

The reference voltage V_r of this converter is connected to the inverting inputs of $2^n - 1$ comparators through the resistor voltage divider. In general, $2^n - 1$ comparators are required for conversion and recording of unknown voltage V_{in} into a $n-$bit word in binary code.

Non-inverting inputs of the comparators are connected to the analogue voltage V_{in} from the S/H circuit. Since all the divider resistors are equal, the voltage at the lowest comparator is $\frac{1}{2^n}V_r$. The maximum voltage at the highest comparator is $\frac{2^n - 1}{2^n}V_r$. The voltage V_{in} compared with the fraction of the voltage V_r determines the output. The relation between V_{in} and V_r causes the outputs of comparators to generate the temperature code. The example of the temperature code for $n = 3$ is shown below.

Input	Output
$0.875V_r \le V_{in} < 1.000V_r$	1111111
$0.750V_r \le V_{in} < 0.875\ V_r$	0111111
$0.625V_r \le V_{in} < 0.750\ V_r$	0011111
$0.500V_r \le V_{in} < 0.625\ V_r$	0001111
$0.375V_r \le V_{in} < 0.500\ V_r$	0000111
$0.250V_r \le V_{in} < 0.375\ V_r$	0000011
$0.125V_r \le V_{in} < 0.250\ V_r$	0000001
$V_{in} < 0.125V_r$	0000000

(1.5)

The conversion of the temperature code into the Gray code and natural binary code, for $n = 3$ bits is shown in Fig. 1.12

6	5	4	3	2	1	0		Gray code			Natural binary code		
0	0	0	0	0	0	0	K_0	0	0	0	0	0	0
						1	K_1	0	0	1	0	0	1
	0				1		K_2	0	1	1	0	1	0
				1			K_3	0	1	0	0	1	1
			1				K_4	1	1	0	1	0	0
		1				1	K_5	1	1	1	1	0	1
	1						K_6	1	0	1	1	1	0
1	1	1	1	1	1	1	K_7	1	0	0	1	1	1
								Y_5	Y_4	Y_3	Y_2	Y_1	Y_0

Fig. 1.12 The conversion of temperature code into Gray code and natural binary code

An example of the logic gate network of $3-$bit A/D converter, with the output signal in the Gray code, is shown in Fig 1.13.

The relations between the temperature code at the output of comparators, and the Gray code at the output of converter are as follows

$$\begin{aligned} Y_5 &= K4 \\ Y_4 &= K_2 \oplus K_6 \\ Y_3 &= K_1 \oplus K_3 + K_5 \oplus K_7 = K_1 \oplus K_3 \oplus K_5 \oplus K_7 \end{aligned} \tag{1.6}$$

The K_0 comparator does not take part in the conversion process (see Fig. 1.12), hence it is not included and not shown in Fig. 1.13.

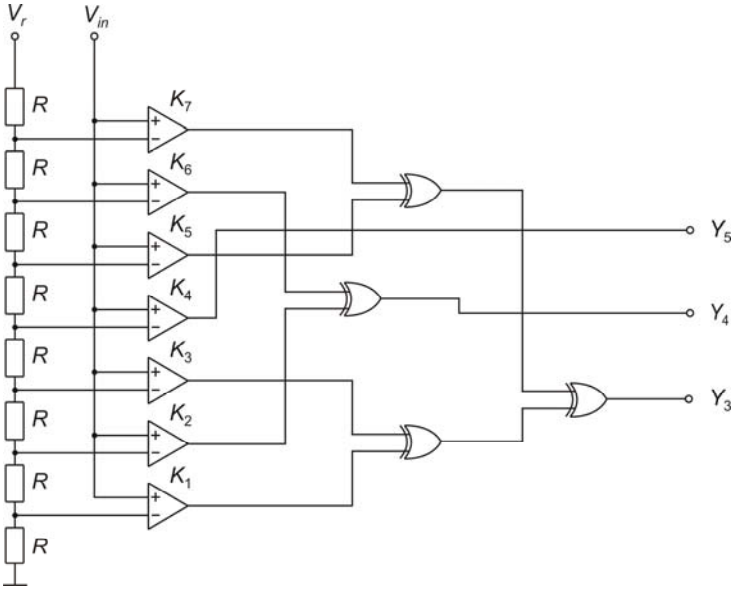

Fig. 1.13 Logic circuit of $3-$bit A/D converter generating output signal in Gray code

1.7.2 A/D Converter with Successive Approximation

The method of successive approximation is based on comparison of the unknown voltage V_{in} with a sequence of precise voltages generated by a controlled D/A converter. There are two basic forms of this type A/D converters, namely with successive approximation and with uniform approximation. The block diagram of a successive approximation method is shown in Fig. 1.14. The corresponding graphs of the clock-generator signal and the voltage under measurement are shown in Fig. 1.15.

In this converter, the register is resetting before a conversion is started (time t_0). As a result, the output voltage of the D/A converter is set to zero. The operation is managed, like any other in the converter, by the control system. The clock generates impulses of voltage V_g, which are fed into the system. Each clock impulse causes voltage V_d to change i.e. the voltage V_d jumps to another value. Each jump of voltage V_d is twice smaller than the previous one. The measuring cycle contains *n* steps of comparison, which are written into the register.

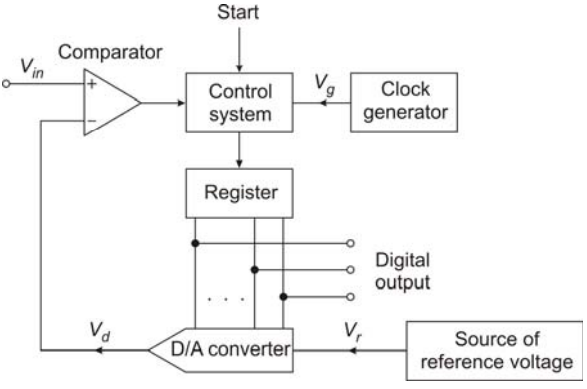

Fig 1.14 A/D converter with successive approximation

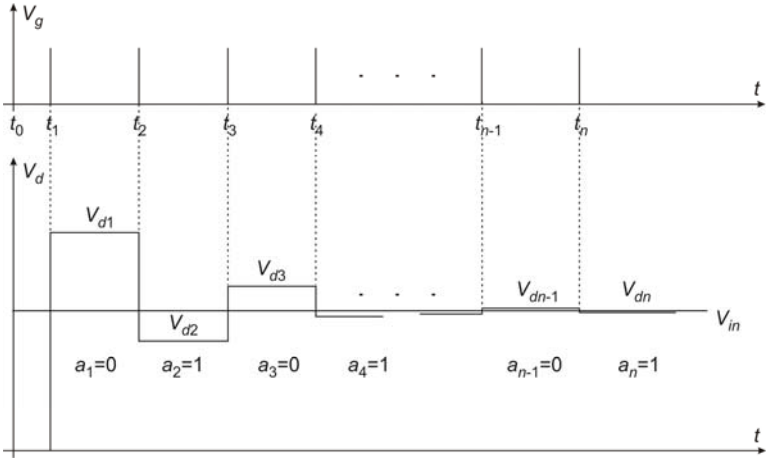

Fig. 1.15 Signal of clock-generator and voltage under measurement

During the first approximation (time t_1), voltage V_{in} is compared with voltage V_{d1}. This voltage is given by

$$V_{d1} = 2^{n-1} V_r \tag{1.7}$$

and V_r denotes the value of the reference voltage related to the least significant bit (LSB).

If a voltage comparison shows $V_{d1} > V_{in}$, which means that the first approximation overestimates V_{in}, then the most significant bit (MSB) is locked at zero. The value $a_1 = 0$ will be recorded accordingly in the register. However, if $V_{d1} < V_{in}$, then the value $a_1 = 1$ will be recorded in MSB of the register.

During the second approximation (time t_2) V_{in} is compared with V_{d2}, where

$$V_{d2} = (2^{n-1} a_1 + 2^{n-2}) V_r \tag{1.8}$$

If $V_{d2} > V_{in}$, then the value $a_2 = 0$ will be recorded in the next in turn bit of the register. However, if $V_{d2} < V_{in}$ then $a_2 = 1$.

For the third approximation (time t_3), V_{d3} is given by

$$V_{d3} = (2^{n-1} a_1 + 2^{n-2} a_2 + 2^{n-3}) V_r \tag{1.9}$$

The value a_3 recorded in the consecutive bit of the register will either be $a_3 = 0$ for $V_{d3} > V_{in}$ or $a_3 = 1$ if $V_{d3} < V_{in}$.

For the $n-th$ approximation (time t_n) V_{dn} is given by

$$V_{dn} = (2^{n-1} a_1 + 2^{n-2} a_2 + ... + 2^1 a_{n-1} + 2^0) V_r \tag{1.10}$$

and $a_n = 0$ if $V_{dn} > V_{in}$ or $a_n = 1$ if $V_{dn} < V_{in}$.

The result of the voltage V_{in} measurement is the binary sequence $a_1, a_2, ..., a_n$ recorded and saved in the register. The full cycle of voltage compensation is relatively short in this converter. It is due to the fact that the jumps of voltage V_d are non-uniform and large during the initial part of the measurement process.

High accuracy of measurements and high speed of response are both the advantages of the converter. However, its complex structure and sensitivity to the external interference and noise are definite disadvantages. In reference to the complexity, the point is that the converter requires high precision voltage dividers.

The converter with uniform compensation, also named staircase-ramp converter, is another type of the successive approximation A/D converter. Fig. 1.16 shows its block diagram and its time-voltage graphs are shown in Fig. 1.17.

In this type of the converter, the voltage V_d is a staircase digital representation, made up of equally increasing steps ΔV_d. Each input signal is equivalent in value to the least significant bit. Clock impulses are fed into the counter. After converting its content into an analogue signal it becomes the voltage V_d.

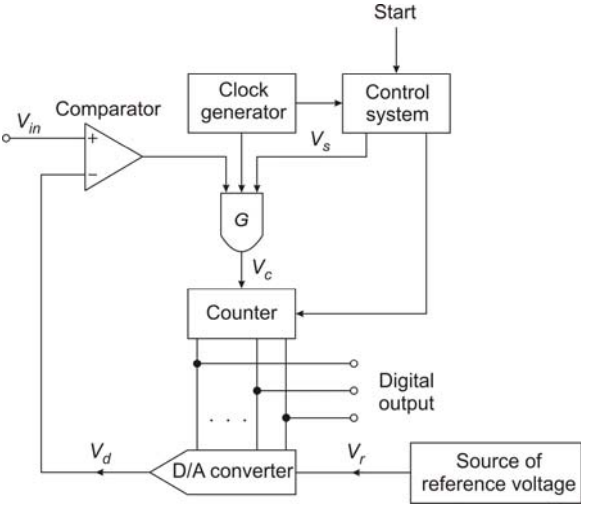

Fig. 1.16 Block diagram of uniform compensation method

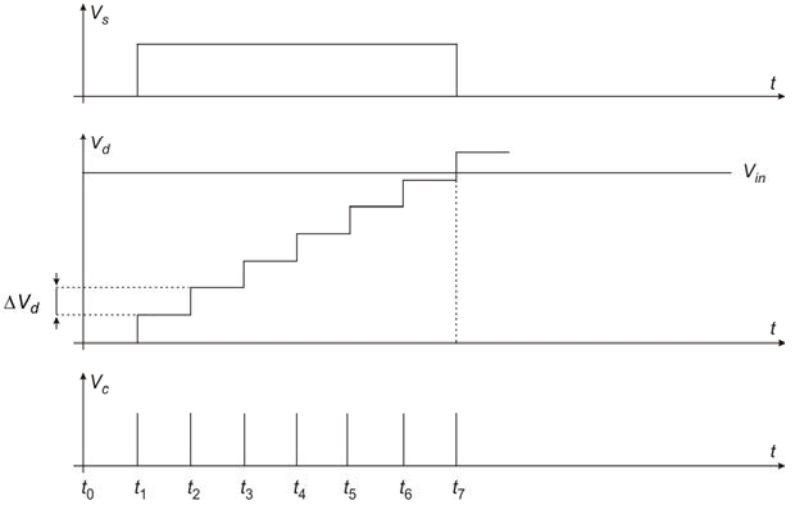

Fig. 1.17 Graphs of time-voltage signals

At first the counter is resetting before a conversion is started. It is controlled by the impulse from the control system. Then the control system starts the count after passing the logic 1 (high level of V_s) into the gate G. The count continues until the generated staircase ramp of the voltage V_d exceeds the measured voltage V_{in}. At this moment the comparator goes to logic zero, the gate G is closed, and it stops the count. The counter output is, at this time, the digital equivalent of the voltage V_{in}.

The time of conversion varies and depends on the value of measured voltage. This is a principal disadvantage of the converter with uniform compensation. For this reason, a modified version of this converter is more often in use, with the reversible counter included into the structure. Such a solution is a significant change in the operation of the converter. The sense of it lies in the fact that each new measuring cycle does not start from zero. After reaching the value of measured voltage, the compensating voltage V_d is tracking further changes of the voltage under measurement, or in other words operates in the follow-up mode. The reverse counter *counts up* for V_{c+} and *counts down* for V_{c-}.

The converter with the reversible counter is often called a follow-up converter. The block diagram of the follow-up A/D converter is shown in Fig. 1.18.

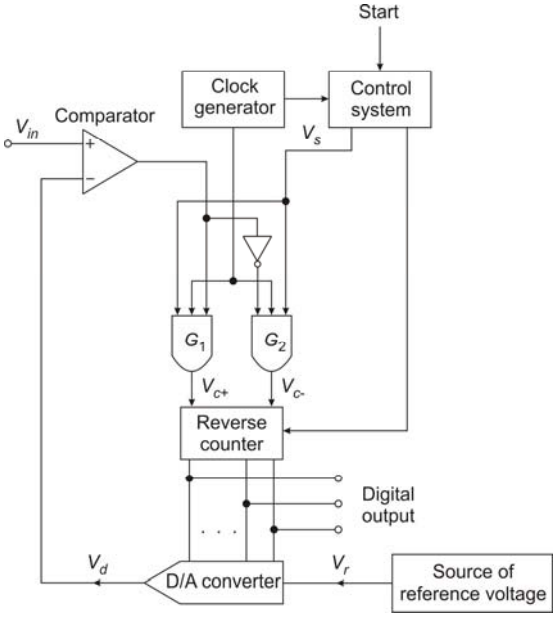

Fig. 1.18 Block diagram of the follow-up A/D converter

1.7.3 Integrating A/D Converters

Analogue-to-digital conversion by integration is based on counting clock impulses, which means the conversion is done indirectly. The operation is completed in two steps. During the first step, the measured voltage is converted into the frequency or time. In the second step, a counting of clock impulses is carried out.

Integrating A/D converters can be divided in two main groups. There are converters with the single integration, in which frequency is used as indirect quantity and converters with the multiple integration, in which time is used as indirect quantity. The block diagram of the A/D converter with the single integration is shown in Fig. 1.19 while Fig. 1.20 shows graphs of the reference voltage and measuring signals at various points of this converter.

There are two main blocks in the A/D converter with the single integration. In the first block, a converter expresses the voltage in terms of the frequency of impulses. The second block is the digital meter of frequency. The main control system controls the whole operation and directly controls start/stop action of individual elements. It determines the conversion cycles of the measured voltage V_{in}.

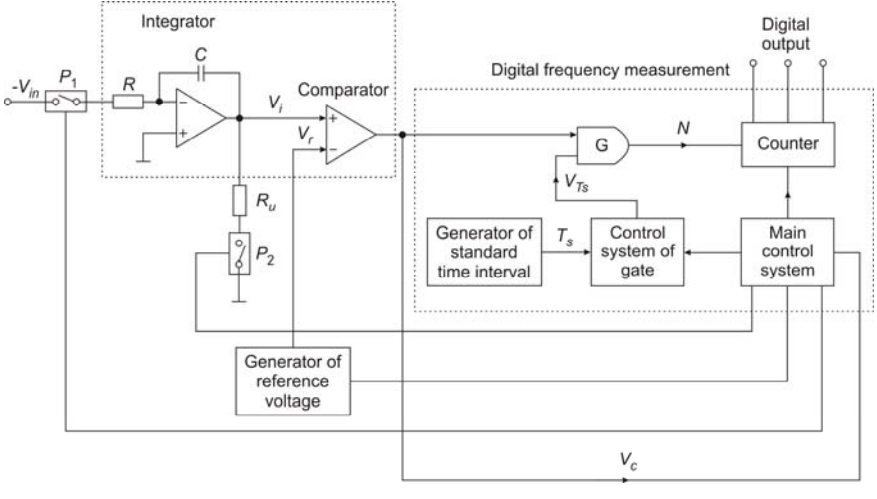

Fig. 1.19 Block diagram of A/D converter with single integration

At the beginning of each measuring cycle, the counter is reset to zero. The generator of reference voltage is turned on and switches on the voltage V_r to the comparator. Then the switch P_1 is closed, and the current proportional to the voltage V_{in} charges a capacitor until the comparator indicates equality of the voltage V_i and V_r. The moment is denoted by t_{in} and at this time the main

control system opens P_1 and closes P_2. The capacitor is discharged through the resistor R_u and the discharge time is denoted t_u. When the time t_u is over, the main control system opens the switch P_2 and closes again the switch P_1. The last action means the start of a new measuring cycle.

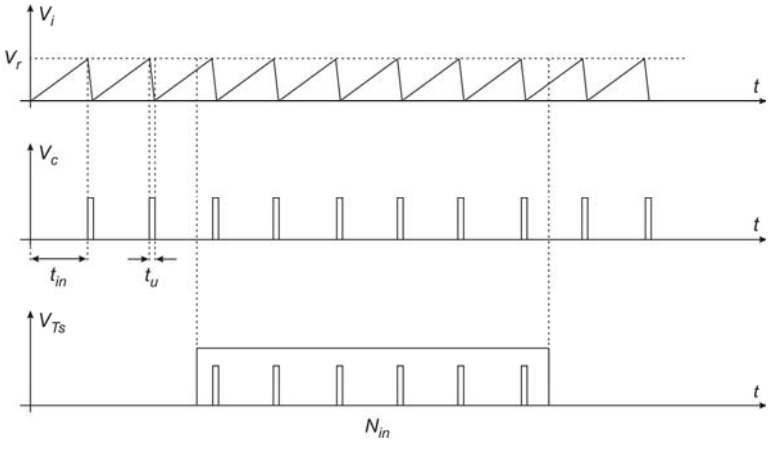

Fig. 1.20 Graphs of time-voltage signals at various points of A/D converter with single integration

The output voltage of the integrator is given by

$$V_i = \frac{1}{RC} V_{in} t \tag{1.11}$$

hence, after time t_{in} when $V_i = V_r$, the Eq. (1.11) changes into

$$V_i = V_r = \frac{1}{RC} V_{in} t_{in} \tag{1.12}$$

Rearranging the equation yields

$$t_{in} = \frac{V_r RC}{V_{in}} \tag{1.13}$$

Assuming that t_u is very small in comparison to t_{in} and can be neglected, the frequency of the discharge impulses is

$$f_{in} = \frac{1}{t_{in}} = \frac{1}{V_r RC} V_{in} \tag{1.14}$$

Denoting by

$$k_f = \frac{1}{V_r RC} \tag{1.15}$$

the frequency f_{in} equals finally

$$f_{in} = k_f V_{in} \tag{1.16}$$

From the Eq. (1.16) it can be seen that the voltage V_{in} to be converted is proportional to f_{in}. The frequency is measured by the frequency digital meter. The measurement is carried out while the gate *G* is open. The gate *G* is controlled by the voltage signal V_{Ts}. The voltage V_{Ts} in turn, and its duration, are both generated and controlled by the generator of standard time interval. The measurement is carried out through the counting of impulses N_{in}

$$N_{in} = T_s f_{in} = T_s k_f V_{in} \tag{1.17}$$

Substituting Eq. (1.15) into Eq. (1.17) and rearranging, the value of measured voltage is

$$V_{in} = \frac{N_{in}}{k_f T_s} = \frac{V_r RC}{T_s} N_{in} \tag{1.18}$$

The A/D converter with the single integration presented above is not often in use. It is due to the non-linearity in the first part of the integrator characteristic. However, its principle of operation is widely applied to structures of other converters, like A/D converters with the multiple integration or sigma delta A/D converters.

Fig. 1.21 shows the block diagram of the A/D converters with the double integration.

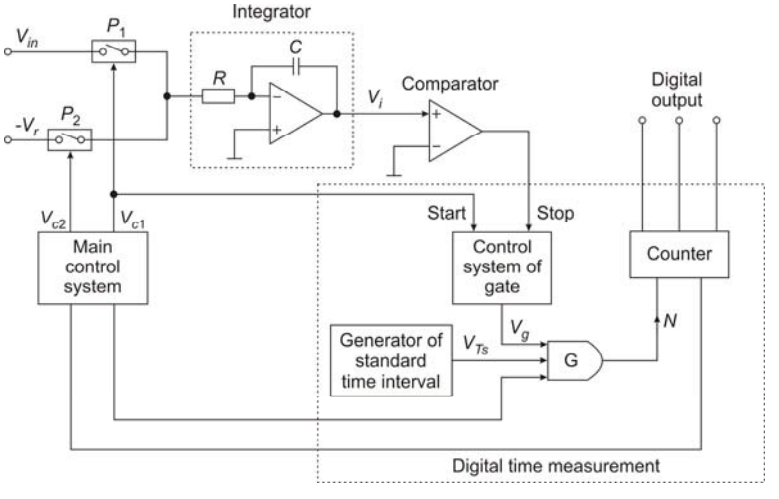

Fig. 1.21 Block diagram of A/D converter with double integration

Graphs of the reference voltage and measuring signals, taken at selected points of the block diagram, are shown in Fig. 1.22.

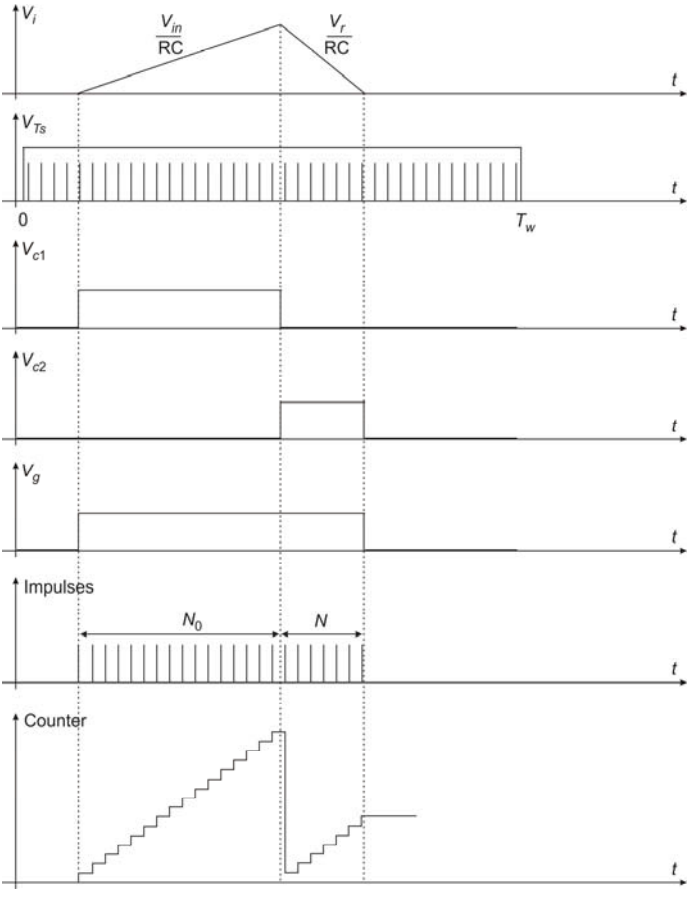

Fig. 1.22 Graphs of time-voltage signals at various points of A/D converter with double integration

The A/D converter with the double integration expresses the value of measured voltage in terms of clock impulses. The conversion, carried out by the A/D converter shown in Fig. 1.21, is completed in two steps. These are defined by the on-off state of the switches P_1 and P_2. When the switch P_1 is closed, the measured voltage V_{in} is applied to the integrator. When the switch P_2 is closed, the input to the integrator is switched from V_{in} to a reference voltage V_r of opposite polarity.

The whole operation is controlled by the main control system. At the beginning of the measuring cycle, the counter is reset to zero. Then the switch P_1 is closed

and integration of the voltage V_{in} begins. The gate *G* is open. The generator of standard time interval produces an impulses sequence during time T_s. Through the open gate *G*, the impulses are passed on to the counter, and a counting starts. It goes on until the counter is overflowed at N_0 impulses**.** At this moment the system turns off the switch P_1 and turns on P_2, resetting the counter at the same time. Counting of impulses starts all over again. The generator of standard time interval produces the impulses sequence and the counter counts down the impulses starting from zero. The output voltage V_i of the integrator decreases down to zero. At this point the gate *G* is closed and the number of impulses counted down is *N*.

The first step of conversion is completed during the time T_{in}

$$T_{in} = N_0 T_w \tag{1.19}$$

where T_w is duration of the impulses.

The output voltage of the integrator during the first step is expressed by

$$V_{i1} = \frac{1}{RC}\int_0^{T_{in}} V_{in}dt = \frac{1}{RC}T_{in}V_{in} = \frac{N_0 T_w}{RC}V_{in} \tag{1.20}$$

In the second step of conversion cycle, integration is completed at the time T_r

$$T_r = NT_w \tag{1.21}$$

and the output voltage of the integrator at this time is given by

$$V_{i2} = \frac{1}{RC}\int_0^{T_r} -V_0 dt = \frac{-1}{RC}T_r V_0 = \frac{-NT_w}{RC}V_r \tag{1.22}$$

The output voltage of the integrator equals zero at the end of the second step. Hence, the sum of V_{i1} and V_{i2} equals zero

$$V_i = \frac{N_0 T_w}{RC}V_{in} - \frac{NT_w}{RC}V_r = 0 \tag{1.23}$$

After rearrangement and simplification we have

$$N = \frac{N_0}{V_r}V_{in} \tag{1.24}$$

The number of impulses N counted down by the counter is proportional to the voltage V_{in}. It also depends on reference voltage V_r and the value of N_0.

Rearranging and substituting Eq. (1.19), (1.21) and (1.24) we finally obtain

$$V_{in} = \frac{T_r}{T_{in}}V_r \tag{1.25}$$

1.7.4 Sigma Delta A/D Converter

A sigma delta converter belongs to the group of converters with the frequency conversion. A classical delta modulator and an adder system are within its structure. The most important advantage of this type of converters is high resolution, up to 24 bits.

Fig. 1.23 shows the block diagram of the sigma delta converter.

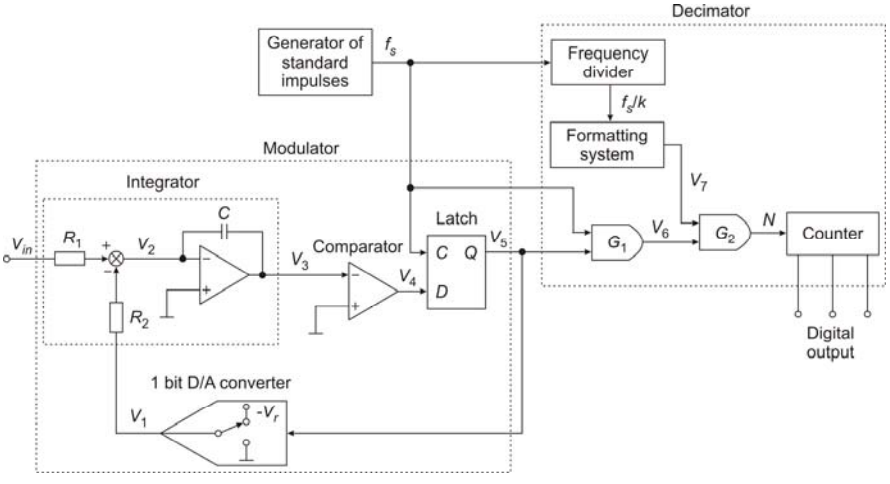

Fig. 1.23 Block diagram of sigma delta converter

Graphs of measuring signals, taken at selected points of this converter, are shown in Fig. 1.24.

Within the sigma delta modulator, the measured voltage V_{in} is added to the output voltage V_1 of the one – bit D/A converter. The summation is done by a summer and the resulting voltage is denoted V_2. The next block is the integrator, in which the voltage V_2 is integrated. At the output of the integrator voltage V_3 has a shape of saw-toothed curve. It is, in turn, changed into the impulse sequence V_4 by the comparator. The number the output impulses from the comparator is in the direct relation to the value of the converted voltage V_{in}.

The voltage V_4 is switched on to the input of the D latch. The latch is synchronized and controlled by the generator of standard impulses connected to the input C, and a impulse sequence V_5 appears at the latch output. At the same time V_5 is the input signal to the D/A converter.

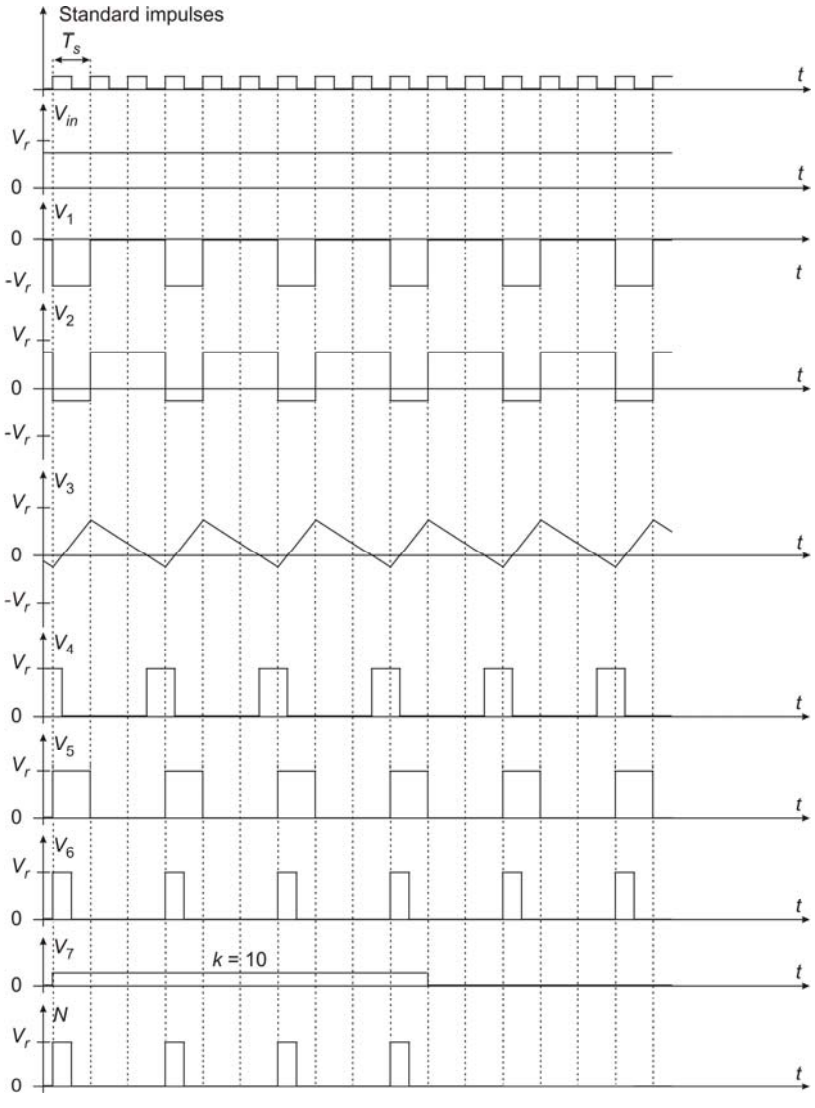

Fig. 1.24 Graphs of time-voltage signals at various points of sigma delta converter

The modulator and the decimator are two main systems within the structure of the sigma delta converter. The decimator changes the serial flow of impulses V_5 into parallel sequences. The first element of the decimator is the gate G_1. The high-level output signal from the generator of standard impulses enables the transmission of V_5 through the gate.

The G_1 output signal is the impulse sequence denoted V_6. The impulse duration of V_6 is only a half of the duration of V_5. The second element of the decimator is the gate G_2. The input signals of the gate G_2 are V_6 and V_7, with V_6 discussed above and V_7 coming as the output signal from the formatting system block.

The frequency f_s of signals from the generator of standard impulses is divided by the number k in the frequency divider, where $k \in N$. The divided frequency f_s / k is processed by the formatting system block and the result is the output signal V_7. The output signal of the gate G_2 is the impulse sequence N, which is counted down by the counter. The period T_c during which the impulses N are counted, equals half of the period of the signal V_7 and is

$$T_c = kT_s \tag{1.26}$$

and T_s is the period of signals from the generator of standard impulses.

For the period T_c, it can be shown that

$$\frac{1}{R_1}\int_0^{T_c} V_{in}dt - N\frac{1}{R_2}\int_0^{T_s} V_r dt = 0 \tag{1.27}$$

The Eq. (1.27) expresses the final result of charging and discharging the capacitor *C* during this time, i.e. the total charge being zero.

Rearranging Eq. (1.27) yields

$$V_{in} = \frac{R_1 V_r T_s}{R_2 T_c} N \tag{1.28}$$

After including Eq. (1.26), the voltage V_{in} is

$$V_{in} = \frac{R_1 V_r}{R_2 k} N \tag{1.29}$$

Eq. (1.29) indicates that the voltage under conversion is proportional to the number of impulses N.

1.8 Input Register

Registers are used to store and manipulate the information data. They store bits of information and, upon an external command, shift those bits one position right or left. The time of storing corresponds to the conversion time of D/A converter. This way registers fulfill the role similar to the sample-and-hold circuits cooperating with A/D converters. Registers are classified according to the method of storing and retrieving information bits. In a serial register, bits are stored or retrieved one

at a time. In a parallel register, all bits of the word are simultaneously stored or retrieved.

Fig. 1.25 presents exemplary logic circuit of parallel register.

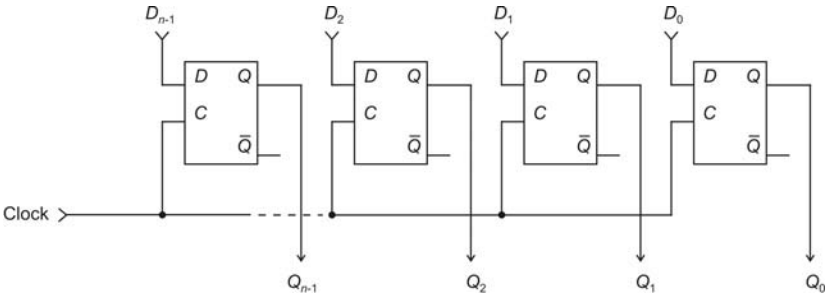

Fig. 1.25 Logic circuit of parallel register

1.9 Digital-to-Analogue Conversion

A D/A conversion is the process of converting input voltage impulses, coming from the output of the DSP, to an analogue voltage. The example of $n-$bit binary-weighted D/A converter is shown in Fig. 1.26 while the example of $n-$bit $R/2R$ ladder D/A converter is presented in Fig. 1.27.

The binary-weighted D/A converter uses a resistor network with resistance values that represent the binary weights of the binary code. The resistor connected to the MSB has a value of R. Each lower-order bit is connected to the resistor which is higher by power of 2. The analogue output is obtained at the junction of the binary weighted resistors. In this type of D/A converter, a number of different value resistors is its disadvantage. For example, the $8-$bit converter uses eight different resistors. If MSB a_0 is connected to R, LSB a_7 is connected to $128R$.

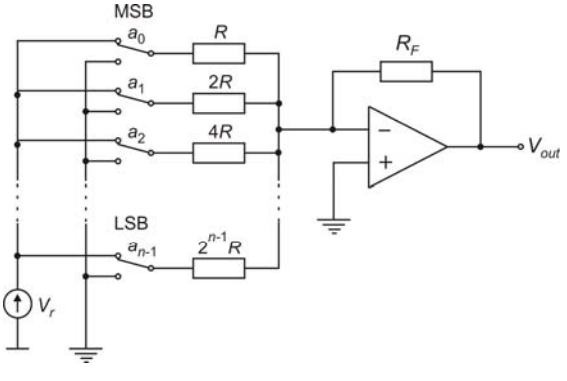

Fig. 1.26 $n-$bit binary-weighted D/A converter

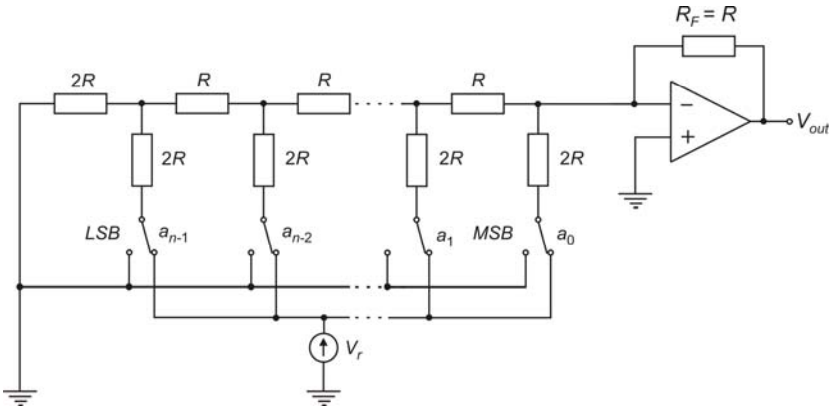

Fig. 1.27 $n-$bit binary-ladder D/A converter

The alternative method of D/A conversion is the $R/2R$ ladder network. It contains two types of resistors only**,** regardless of the number of bits, and one resistor is twice as large as the other. The value of ladder resistors connected to register bits is $2R$, and the value of resistors connected between nodes is R. It is easy to check that the resistance, looking from any node towards terminating resistor, is $2R$. The output voltage V_{out} equals to

$$V_{out} = k_u V_r (a_1 2^{-1} + a_2 2^{-2} + ... + a_n 2^{-n}) \tag{1.30}$$

and V_r is given by

$$V_r = 2^n \,[\mathrm{V}] \tag{1.31}$$

were in (1.30) k_u is amplification of operational amplifier.

1.10 Reconstruction Filter

Fig. 1.28 shows the signal obtained at the output of D/A converter.

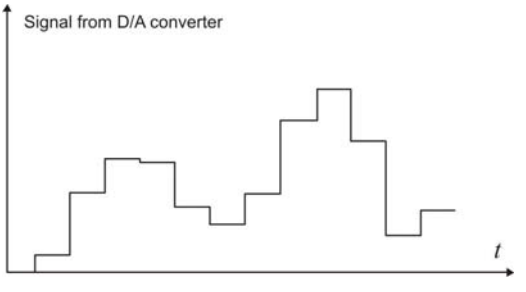

Fig. 1.28 Output signal from D/A converter

The reconstruction filter is used for smoothing and tapering of the stairs waveform. In result, the analogue signal is obtained. Fig 1.29 shows the signal from D/A converter after reconstruction.

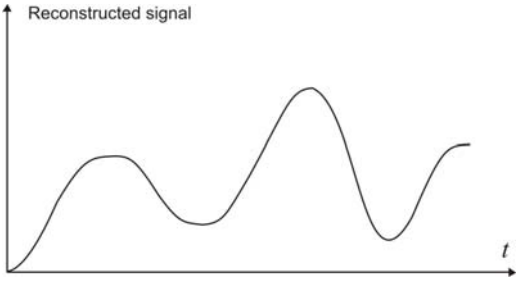

Fig. 1.29 Output signal from A/D converter after reconstruction

1.11 DSP

Digital signal processor can read, write and manipulate digital signals only. The signals converted into the digital form are stored within DSP as binary numbers, usually in the form of combination of 8, 16 or 32 bits. DSP can perform various operations on the incoming data such as removing unwanted interference, increasing some signal frequencies and reducing others, detecting and correcting errors in transmitted codes etc. Its task is to handle data according to the assumed calculation algorithms.

The successive samples of signals are processed using the algorithms with the appropriately selected mathematical operations and with the use of digital filters. The application of the digital filters also enables a change of the DSP setup, followed by an optional change of its frequency characteristics.

Two types of digital filters can be distinguished:

- Finite Impulse Response, abbr. FIR
- Infinite Impulse Response, abbr. IIR.

There are some important differences between these filters. In the case of FIR, the calculations related to a consecutive sample are based on the samples, which have earlier been digitally filtered, and the current sample. The number of the samples filtered earlier and taken into calculations depends on the filter grade. In the case of IIR, all samples filtered earlier are taken into account. The FIR filters are the elements having good stability. Due to this, it can be assumed that their phase characteristic is exactly linear. FIR filters can be used for design and all applications where such a linear characteristic is required or a full control of the system phase response is recommended.

Properties of IIR filters are different. Their phase response is much worse than in the case of FIR filters. It is due to some non-linearity at the edges of bands. However, IIR filters are much faster in applications of calculation algorithms.

Another important application of DSP is the fast spectroanalysis FFT.

1. 12 Control System

The control system has several tasks. These include:

- generation of the start signal for sample-and-hold
- generation of the start signal for A/D conversion
- control of address inputs of multiplexers/demultiplexers
- control of register and D/A converter.

References

[1] Bentley, J.P.: Principles of measurement systems. Pearson Pentice Hall, New York (1996)
[2] Doebelin, E.O.: Measurement Systems. Applications and Design. McGraw-Hill, Boston (2004)
[3] Floyd, T.L.: Digital fundamentals with PLD programming. Pearson Prentice Hall (2006)
[4] Hoeschele, D.F.: Analog-to-digital and digital-to-analog conversion techniques. Wiley, New York (1994)
[5] Johnson, E., Karim, M.: Digital Design. PWS-KENT Publishing Company, Boston (1987)
[6] Nawrocki, W.: Measurement systems and sensors. Artech House, London (2005)
[7] Norman, R.: Principles of Bioinstrumentation. John Wiley & Sons, New York (1988)
[8] Stabrowski, M.: Cyfrowe przyrzady pomiarowe. PWN, Warszawa (2002)
[9] Stranneby, D.: Digital signal processing: DSP and applications. Linaere House (2001)
[10] Tumanski, S.: Technika pomiarowa. WNT, Warszawa (2007)

Chapter 2
Sensors

2.1 Strain Gauge Sensors

Strain gauge sensors are the fundamental sensing elements for many types of sensors e.g. force sensors, torque sensors, pressure sensors, acceleration sensors, etc. They are applied to measure strain. Having strain measured and using Young's modulus E and geometric sizes, stress can be calculated. Finally, from these calculations an unknown and investigated quantity can be found, which is applied and acts on an object under test. Strain gauge principal of operation takes advantage of the physical property of the variety of changes of electrical resistance resulting from its elongation or shortening. If a strip of conductive material is stretched, it becomes skinnier and longer resulting in an increase of its resistance R, while if it is compressed, it becomes shorten and broaden resulting in decrease of its resistance. The principle of operation of a common metallic strain gauges is based on a change of a conductor resistance. Let us present the resistance of electrical conductor in the following form

$$R = \rho \frac{l}{S} \tag{2.1}$$

hence, a relative increment of R equals

$$\frac{\Delta R}{R} = \frac{\rho}{S}\frac{\Delta l}{l} \quad \text{and} \quad \frac{\Delta l}{l} >> \frac{\Delta S}{S} \tag{2.2}$$

where R, ρ, l, S are resistance, resistivity, length and cross section of an electrical conductor, respectively.

Strain gauges are arranged in a wide choice of shapes and sizes depending on variety of application.

Most often however, they are made as a long, thin conductive strip in a zigzag pattern of parallel lines.

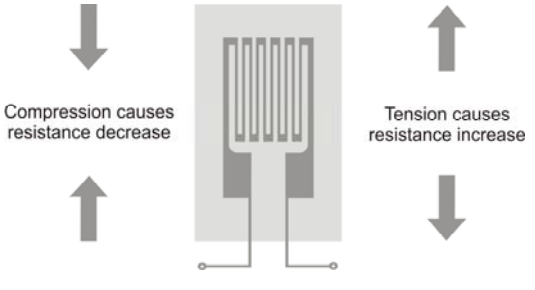

Fig. 2.1 Strain gauge

The strain gauges are glued, usually to a much larger object under test, by means of a special glue. Through the glue, the strain is transferred from the object under test to the strain gauge. In this way, the strain ε of object under test is equal to the strain of gauges. Therefore, for the strain gauge we can write

$$\varepsilon_R = k\,\varepsilon \tag{2.3}$$

where

$\varepsilon_R = \dfrac{\Delta R}{R}$ is the relative increment of strain gauge resistance, the strain $\varepsilon = \dfrac{\Delta l}{l}$ is the relative increment of strain gauge length, k – strain gauge constant.

In the measuring circuit, strain gauges work in a full bridge configuration with a combination of four active gauges shown in Fig. 2.2a, or in half a bridge with two active gauges. In this second case, half a bridge is completed with two precision resistors R_3 and R_4 – Fig. 2.2b.

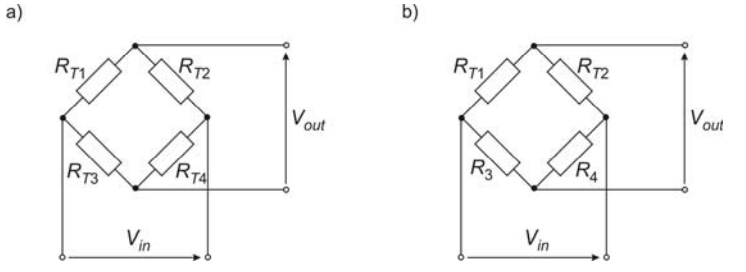

Fig. 2.2 A strain gauge bridge circuits: a) full bridge, b) half a bridge

Assume that the object under test is subjected to a strain. Let us determine increment ΔV_{out} of the output voltage of the bridge effected by increment of all strain gauges resistors $R_{Ti} \Rightarrow R_{Ti} + \Delta R_{Ti}$, $i = 1 \div 4$. For the sake of simplification, let us assume that the strain gauges are arranged in a full bridge configuration, the source V_{in} supplying the bridge has the internal resistance equals zero and the output denoted V_{out} is unloaded as it is connected to the amplifier of infinite resistance. So we have

$$\frac{V_{out}}{V_{in}} = \frac{R_{T1}R_{T4} - R_{T2}R_{T3}}{(R_{T1} + R_{T2})(R_{T3} + R_{T4})} \tag{2.4}$$

and

$$\frac{\Delta V_{out}}{V_{in}} = \frac{(R_{T1} + \Delta R_{T1})(R_{T4} + \Delta R_{T4}) + \ldots}{(R_{T1} + \Delta R_{T1})(R_{T3} + \Delta R_{T3}) + (R_{T2} + \Delta R_{T2})(R_{T3} + \Delta R_{T3}) + \ldots} \tag{2.5}$$

$$\frac{\ldots - (R_{T2} + \Delta R_{T2})(R_{T3} + \Delta R_{T3})}{\ldots + (R_{T1} + \Delta R_{T1})(R_{T4} + \Delta R_{T4}) + (R_{T2} + \Delta R_{T2})(R_{T4} + \Delta R_{T4})}$$

After easy rearrangement, we get

$$\frac{\Delta V_{out}}{V_{in}} = \frac{\varepsilon_{R1} + \varepsilon_{R4} - \varepsilon_{R2} - \varepsilon_{R3}}{4 + 2\left(\frac{\Delta R_{T1}}{R_{T1}} + \frac{\Delta R_{T3}}{R_{T3}} + \frac{\Delta R_{T2}}{R_{T2}} + \frac{\Delta R_{T4}}{R_{T4}}\right)} \tag{2.6}$$

The increment of resistance in elastic limits of the gauge material and of the object under test may change only a fraction of a percent. Because of it, we can assume that

$$2\left(\frac{\Delta R_{T1}}{R_{T1}} + \frac{\Delta R_{T3}}{R_{T3}} + \frac{\Delta R_{T2}}{R_{T2}} + \frac{\Delta R_{T4}}{R_{T4}}\right) << 4 \tag{2.7}$$

and finally Eq. (2.6) can be simplified to the form

$$\frac{\Delta V_{out}}{V_{in}} = \frac{1}{4}(\varepsilon_{R1} + \varepsilon_{R4} - \varepsilon_{R2} - \varepsilon_{R3}) \tag{2.8}$$

or

$$\frac{\Delta V_{out}}{V_{in}} = \frac{k}{4}(\varepsilon_1 + \varepsilon_4 - \varepsilon_2 - \varepsilon_3) \tag{2.9}$$

where strain gauge constant $k \approx 2$.

2.1.1 Temperature Compensation

Strain gauges should be glued in to an object under test, and connected in a bridge circuit, in a special way indicated by Eq. (2.9). If the strains ε_1 and ε_4, related to the gauges R_{T1} and R_{T4} of the bridge shown in Fig. 2.2a, are positive, then the strains ε_2 and ε_3 of the gauges R_{T2} and R_{T3} should be negative. This way the strains add together and the output voltage has a maximum value. At the same time, such a connection makes possible the compensation of thermal effect. The temperature effect causes a change of strain in each of the strain gauges involved. The change denoted $+\varepsilon_T$ is due to thermal expansion of the object under test. Including this effect into Eq. (2.9), we can write

$$\begin{aligned}\frac{\Delta V_{out}}{V_{in}} &= \frac{k}{4}[(\varepsilon_1 + \varepsilon_T) + (\varepsilon_4 + \varepsilon_T) - (\varepsilon_2 + \varepsilon_T) - (\varepsilon_3 + \varepsilon_T)] \\ &= \frac{k}{4}(\varepsilon_1 + \varepsilon_4 - \varepsilon_2 - \varepsilon_3)\end{aligned} \tag{2.10}$$

Examining the Eq. (2.10), it is easy to notice that the influence of temperature in such a circuit is compensated. The same reasoning can be applied to the half a bridge circuit and the connection of strain gauges into it. They should be connected e.g. in the branch R_{T1} for $+\varepsilon_1$ and the branch R_{T2} for $-\varepsilon_2$. It renders certain the temperature compensation because

$$\frac{\Delta V_{out}}{V_{in}} = \frac{k}{4}[(\varepsilon_1 + \varepsilon_T) - (\varepsilon_2 + \varepsilon_T)] = \frac{k}{4}(\varepsilon_1 - \varepsilon_2) \tag{2.11}$$

However, the sensitivity of the arrangement is twice lower. If the temperature compensation is not possible through the appropriate arrangement and connection of active strain gauges, dummy gauges are applied. In Figs. 2.3 – 2.7 the diagram shows the force F and its components F_x and F_y acting on a beam. The beam is bent as the result of action. The aim is to measure the forces using strain gauges in various configurations, and include the temperature compensation for the bent beam.

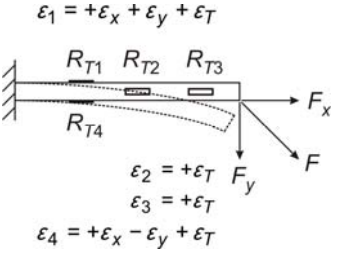

Fig. 2.3 Measurement of the component F_x using full bridge R_{T2}, R_{T3} – dummy gauges

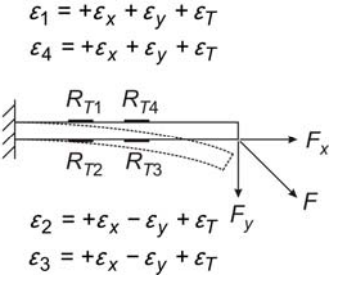

Fig. 2.4 Measurement of the component F_y using full bridge

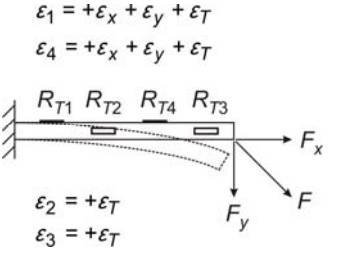

Fig. 2.5 Concurrent measurement of the components F_x and F_y using full bridge R_{T2}, R_{T3} – dummy gauges

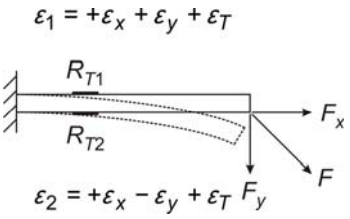

Fig. 2.6 Measurement of the component F_y using half a bridge

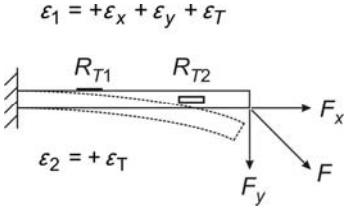

Fig. 2.7 Concurrent measurement of the components F_x and F_y using half a bridge R_{T2} – dummy gauge

2.1.2 Lead Wires Effect

Lead wires are part of a gauge installation. Their resistance may have an important influence during measurements with the use of strain gauges. The voltage drop due to this resistance could impair the performance and decrease sensitivity of the measuring strain gauge system. Hence, the resistance should always be taken into account, particularly in case of longer lead wires.

We shall now consider the setup shown in Fig. 2.8. Let us assume that all the strain gauges connected in the bridge have the same resistance, the fact that should always be a good practice, i.e. $R_{Ti} = R_T$ for $i = 1 \div 4$. Power is supplied to the bridge by the lead wires of the resistance r. It is easy to derive the expression for the voltage $V_{in}^{'}$ connected directly to the bridge and to note that it is smaller than the voltage V_{in} across the lead wire terminals

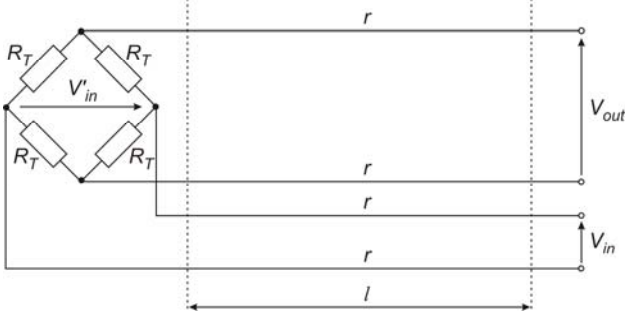

Fig. 2.8 Full bridge and lead wires

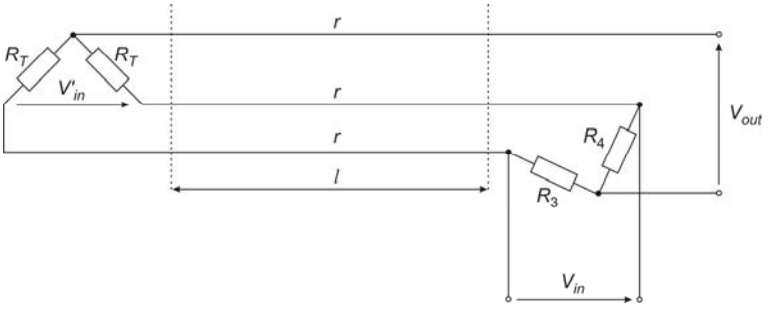

Fig. 2.9 Half a bridge and lead wires

$$V_{in}^{'} = V_{in}\frac{R_T}{R_T + 2r} \tag{2.12}$$

For the half a bridge shown in Fig. 2.9, such a voltage is given by

$$V_{in}^{'} = V_{in}\frac{R_T}{R_T + r} \tag{2.13}$$

Substituting the voltage $V_{in}^{'}$ instead of V_{in} into Eq. (2.10) and (2.11), we get

$$\frac{\Delta V_{out}}{V_{in}} = \frac{k}{4}\frac{R_T}{R_T + 2r}(\varepsilon_1 + \varepsilon_4 - \varepsilon_2 - \varepsilon_3) \tag{2.14}$$

for the full bridge, and

$$\frac{\Delta V_{out}}{V_{in}} = \frac{k}{4}\frac{R_T}{R_T + r}(\varepsilon_1 - \varepsilon_2) \tag{2.15}$$

for half a bridge.

Let us examine the Eq. (2.14) and (2.15). If the resistance $2r$ in the case of (2.14), and the resistance r in (2.15), are equal to the resistance of the strain gauge R_T, then the sensitivity of the setup drops by half.

2.1.3 Force Measurement

During force measurements, for uniaxial stresses the following relations hold

$$\varepsilon = \frac{\sigma}{E} \quad \text{and} \quad \sigma = \frac{F}{S} \tag{2.16}$$

where σ is the stress, E is Young's modulus for steel, S is a cross-sectional area and F is a force applied to the object under test.

2.1.4 Torque Measurement

Torsional moment of the shaft can be measured directly by means of the appropriate location of strain gauges. The gauges are glued in along the main stress axes, where the strains have opposite signs.

Fig. 2.10 shows how the strain gauges are glued to the surface for the measurement of torsional moment.

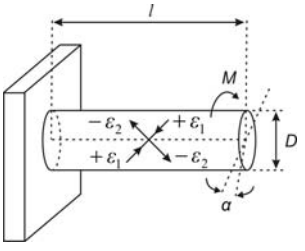

Fig. 2.10 Location of strain gauges for the torque measurement

For torque measurements of shafts, the following relations hold

$$\varepsilon_1 = -\varepsilon_2 = \frac{8M}{\pi G D^3} = \frac{D}{4l}\alpha \quad \text{and} \quad G = \frac{E}{2(1+\nu)} \tag{2.17}$$

while for a tube

$$\varepsilon_1 = -\varepsilon_2 = \frac{8MD}{\pi G(D^4 - d^4)} \tag{2.18}$$

where G is Kirchhoff's modulus, Poisson's ratio for steel is $\nu \approx 0.3$, M is the torque, α is the angle of shaft torsion, the shaft diameter is D and its length l, and d is the inside diameter of tube.

2.1.5 Pressure Measurement

Steel diaphragms with pressure gauges glued in on them can be used for pressure measurement. Fig. 2.11 shows the diaphragm pressure gauges. The circular diaphragm fixed in the enclosure is shown in Fig. 2.12. The extended connector pipe of the enclosure is screwed in into a pressure conduit, in which the pressure is to be measured.

Pressure to be measured causes a deflection of steel diaphragm, which leads to development of stresses in it. During the pressure measurement, a radial stress and, perpendicular to it, a tangential stress are both utilized.

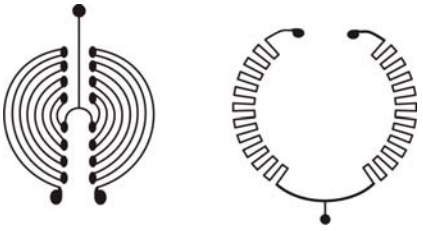

Fig. 2.11 Diaphragms pressure gauges

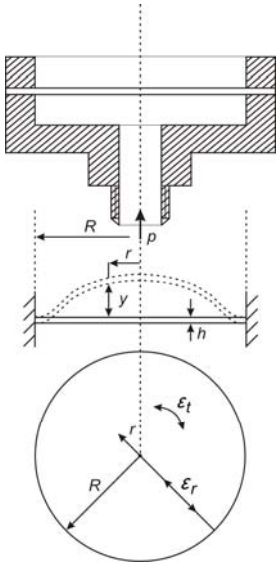

Fig. 2.12 Circular diaphragm fixed in the enclosure

Radial stress σ_r equals

$$\sigma_r = \frac{3}{8} p \left(\frac{R}{h}\right)^2 \left[(1+\nu) - (3+\nu)\left(\frac{r}{R}\right)^2 \right] \tag{2.19}$$

while tangential stress σ_t is

$$\sigma_t = \frac{3}{8} p \left(\frac{R}{h}\right)^2 \left[(1+\nu) - (3\nu+1)\left(\frac{r}{R}\right)^2 \right] \tag{2.20}$$

Relations between radial and tangential stresses σ_r, σ_t and radial and tangential strains ε_r, ε_t for biaxial state of stresses are as follows

$$\varepsilon_r = \frac{\sigma_r - \nu\,\sigma_t}{E} \tag{2.21}$$

and

$$\varepsilon_t = \frac{\sigma_t - \nu\,\sigma_r}{E} \tag{2.22}$$

Substituting (2.19) and (2.20) into (2.21) and (2.22) gives finally

$$\varepsilon_r = \frac{3}{8}\,p\,\frac{(1-\nu^2)}{E}\left(\frac{R}{h}\right)^2\left[1-3\left(\frac{r}{R}\right)^2\right] \tag{2.23}$$

and

$$\varepsilon_t = \frac{3}{8}\,p\,\frac{(1-\nu^2)}{E}\left(\frac{R}{h}\right)^2\left[1-\left(\frac{r}{R}\right)^2\right] \tag{2.24}$$

Examination of the Eq. (2.23) and (2.24) makes possible to indicate places in which maximum strain occurs, and where gauges should be glued in. For the strain ε_r, it is the peripheral edge of diaphragm $r = R$. The corresponding expression for $\max \varepsilon_r$ is

$$\max \varepsilon_r = -\frac{3}{4}\,p\,\frac{(1-\nu^2)}{E}\left(\frac{R}{h}\right)^2 \tag{2.25}$$

For the strain ε_t, it is the centre of diaphragm $r = 0,$ and

$$\max \varepsilon_t = \frac{3}{8}\,p\,\frac{(1-\nu^2)}{E}\left(\frac{R}{h}\right)^2 \tag{2.26}$$

It is not allowed to glue gauges to the places where the radial strain ε_r and tangential one ε_t equal zero. These places are for the radial strain

$$\varepsilon_r = 0 \quad \text{for} \quad r = \frac{\sqrt{3}}{3}R \tag{2.27}$$

and for the tangential strain

$$\varepsilon_t = 0 \quad \text{for} \quad r = R \tag{2.28}$$

Fig. 2.13 shows characteristics of the stress σ_r and σ_t as a function of the diaphragm radius r, while Fig. 2.14 shows characteristics of strains ε_r and ε_t.

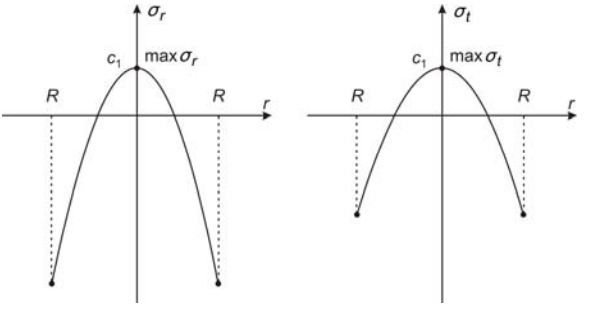

Fig. 2.13 Characteristics σ_r and σ_t, $c_1 = \frac{3}{8} p \left(\frac{R}{h}\right)^2 (1+\nu)$

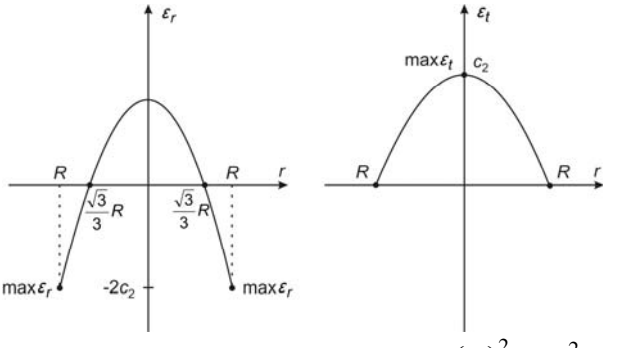

Fig. 2.14 Characteristics ε_r and ε_t, $c_2 = \frac{3}{8} p \left(\frac{R}{h}\right)^2 \frac{(1-\nu^2)}{E}$

2.2 Capacitive Sensors

A capacitive sensor for pressure measurements is based on a capacitor of varying capacity related to measured quantity. Fig. 2.15 shows a capacitive pressure sensor that has a fixed plate and a movable one. The movable plate is a circular flat diaphragm, and the other one is a metal housing. When the pressure is applied to the diaphragm, its motion is a measure of applied pressure. The motion of the diaphragm changes the distance between the diaphragm and the fixed plate. The capacitance of the sensor increases to $C + \Delta C$ and the output of the sensor is the change in capacitance ΔC. The new value of capacitance $C + \Delta C$ is

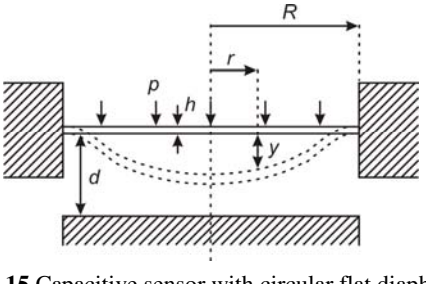

Fig. 2.15 Capacitive sensor with circular flat diaphragm

$$C + \Delta C = \int_0^R dC = \int_0^R \varepsilon_0 \frac{2\pi r}{d - y} dr \tag{2.29}$$

where C is the capacitance before the diaphragm sagging.

Let us derive the relation between the change of capacitance and the pressure to be measured. At first, for the sake of simplification, let us note that for the small value of the ratio y/d, we have

$$\frac{1}{d - y} \approx \frac{1}{d}\left(1 + \frac{y}{d}\right) \quad \text{if} \quad \frac{y}{d} << 1 \tag{2.30}$$

Substituting (2.30) into (2.29), we get

$$C + \Delta C = \frac{2\pi\varepsilon_0}{d} \int_0^R \left(1 + \frac{y}{d}\right) r dr \tag{2.31}$$

The diaphragm sag y at the radius r (Fig. 2.15) is

$$y = \frac{3p(1-\nu)(R^2 - r^2)^2}{16 E h^3} \tag{2.32}$$

Substituting (2.32) into (2.31), we get

$$C + \Delta C = \frac{2\pi\varepsilon_0}{d}\left[\int_0^R r\, dr + \frac{3p(1-\nu)}{16 E h^3 d}\left(\int_0^R R^4 r\, dr - 2\int_0^R R^2 r^3 dr + \int_0^R r^5 dr\right)\right] \tag{2.33}$$

After integration and simplification, we obtain

$$C + \Delta C = \frac{2\pi\varepsilon_0}{d}\left[\frac{R^2}{2} + \frac{p}{2}\frac{(1-\nu)R^6}{16 E h^3 d}\right] \tag{2.34}$$

It can be easily noticed, that the capacitance of the sensor, before the diaphragm sag, is given by the first term of the sum (2.34)

$$C = \frac{\pi R^2 \varepsilon_0}{d} \tag{2.35}$$

hence the absolute value of the capacitance increment is

$$\Delta C = \frac{\pi\varepsilon_0 p}{d^2}\frac{(1-\nu)R^6}{16 E h^3} \tag{2.36}$$

and the relative value is

$$\frac{\Delta C}{C} = p\frac{(1-\nu)R^4}{16 E h^3 d} \tag{2.37}$$

From Eq. (2.37), it can be seen that the relative value of the capacitance increment is directly proportional to the measured pressure.

Usually, the capacitive sensors are components of A.C. bridge dedicated for capacitance measurement, for example the Wien's bridge.

2.3 Inductive Sensors

Inductive sensors are inductive devices for the measurement of small displacements. In inductive sensors, the principle of operation is based on the relations between their magnetic and electric circuits. More specifically, the change of the reluctance of the magnetic circuit leads to the change of the impedance in the electric circuit.

Fig. 2.16 shows the inductive sensor with the magnetic circuit consisting of the fixed iron core, the movable armature and the variable air gap.

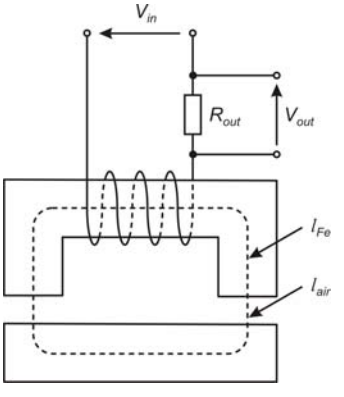

Fig. 2.16 Inductive sensor

The current in the setup shown in Fig. 2.16 is

$$i = \frac{V_{in}}{R + j\omega L} = \frac{V_{in}}{R + j\dfrac{\omega z^2}{R_{\mu Fe} + R_{\mu air}}} \tag{2.38}$$

where L is the circuit inductance

$$L = \frac{z^2}{R_\mu} = \frac{z^2}{R_{\mu Fe} + R_{\mu air}} = \frac{z^2}{\dfrac{l_{Fe}}{\mu_{Fe} S_{Fe}} + \dfrac{l_{air}}{\mu_{air} S_{air}}} \tag{2.39}$$

and R is the equivalent resistance of the winding circuit r_{cu}, R_{out} and R_{Fe}

$$R = r_{cu} + R_{out} + R_{Fe} \tag{2.40}$$

where

$$R_{Fe} = R_h + R_e \tag{2.41}$$

The resistance R_{Fe} is related to iron loss in the core, i.e. the power dissipated as heat, and is a sum of the resistances R_h and R_e. The resistance R_h is related to hysteresis loss

$$P_h = \sigma_h m f B^2 \tag{2.42}$$

and the resistance R_e is related to eddy current loss

$$P_e = \sigma_e m f^2 B^2 \tag{2.43}$$

Symbols and notations used in the Eq. (2.38) till (2.43) have the following meaning:

- μ_{air} and μ_{Fe} are magnetic permeabilities of air and iron
- σ_h and σ_e are loss coefficients for hysteresis and eddy currents
- B is a flux density in the iron core
- l_{air} and l_{Fe} are the lengths of the air gap and the iron core
- S_{air} and S_{Fe} are the cross-sectional areas of the air gap and the iron core
- m is the mass of the iron.

After examination of the Eq. (2.38) and (2.39), it is easy to reach the conclusion that the current in the sensor coil is related to the changes of air gap length. It makes possible to measure the air gap length indirectly, through the measurement of the voltage V_{out} across the resistor R_{out}

$$V_{out} = R_{out} \frac{V_{in}}{\sqrt{R^2 + \left(\frac{\omega z^2}{\frac{l_{Fe}}{\mu_{Fe} S_{Fe}} + \frac{l_{air}}{\mu_{air} S_{air}}} \right)^2}} \tag{2.44}$$

Quite often, inductive sensors are connected into bridge circuits, for instance in the Maxwell or Maxwell-Wien bridges. Such an arrangement generates the problem of the phase shift φ between current and voltage, which for ω = const. should also be constant

$$\varphi = arctg \frac{\omega L}{R} \tag{2.45}$$

It can be achieved on condition that the relative increments of inductance and resistance are equal

$$\frac{\frac{dL}{dl_{air}}}{\frac{L}{}} = \frac{\frac{dR}{dl_{air}}}{R} \tag{2.46}$$

Assuming the identity of the cross-sectional areas $S_{Fe} = S_{air} = S$, let us calculate the derivatives shown in Eq. (2.46). For the left side of the equation, we get

$$\frac{\frac{dL}{dl_{air}}}{L} = \frac{\frac{d}{dl_{air}}\left(\frac{z^2}{\frac{l_{Fe}}{\mu_{Fe}S_{Fe}} + \frac{l_{air}}{\mu_{air}S_{air}}}\right)}{\frac{z^2}{\frac{l_{Fe}}{\mu_{Fe}S_{Fe}} + \frac{l_{air}}{\mu_{air}S_{air}}}} = \frac{-1}{\mu_{air}\left(\frac{l_{Fe}}{\mu_{Fe}} + \frac{l_{air}}{\mu_{air}}\right)} \tag{2.47}$$

In order to calculate the derivative of the right side of Eq. (2.46), the details of R_{Fe} must be considered. The total power loss in this resistance is

$$i^2 R_{Fe} = (\sigma_h + \sigma_e f)\, m\, f\, B^2 \tag{2.48}$$

The equation for the flux density is

$$B = \frac{i\, z}{S R_\mu} \tag{2.49}$$

Substituting (2.49) into (2.48), we get

$$i^2 R_{Fe} = (\sigma_h + \sigma_e f)\, mf \frac{i^2 z^2}{S^2\left(\frac{l_{Fe}}{\mu_{Fe}S} + \frac{l_{air}}{\mu_{air}S}\right)^2} \tag{2.50}$$

Finally

$$R_{Fe} = (\sigma_h + \sigma_e f)\, mf \frac{z^2}{\left(\frac{l_{Fe}}{\mu_{Fe}} + \frac{l_{air}}{\mu_{air}}\right)^2} \tag{2.51}$$

Returning to Eq. (2.46) and substituting Eq. (2.51) into it, we get

$$\frac{\frac{dR}{dl_{air}}}{R} = \frac{\frac{d}{dl_{air}}\left(r_{cu} + R_{out} + (\sigma_h + \sigma_e f)\, mf \frac{z^2}{\left(\frac{l_{Fe}}{\mu_{Fe}} + \frac{l_{air}}{\mu_{air}}\right)^2}\right)}{r_{cu} + R_{out} + (\sigma_h + \sigma_e f)\, mf \frac{z^2}{\left(\frac{l_{Fe}}{\mu_{Fe}} + \frac{l_{air}}{\mu_{air}}\right)^2}} \tag{2.52}$$

$$= \frac{-2R_{Fe}}{\mu_{air}(r_{cu} + R_{out} + R_{Fe})\left(\frac{l_{Fe}}{\mu_{Fe}} + \frac{l_{air}}{\mu_{air}}\right)}$$

The constant phase shift condition (2.46) can be now expressed in the form

$$\frac{1}{\mu_{air}\left(\frac{l_{Fe}}{\mu_{Fe}} + \frac{l_{air}}{\mu_{air}}\right)} = \frac{2R_{Fe}}{\mu_{air}(r_{cu} + R_{out} + R_{Fe})\left(\frac{l_{Fe}}{\mu_{Fe}} + \frac{l_{air}}{\mu_{air}}\right)} \tag{2.53}$$

It is easy to prove that the condition is satisfied, if

$$r_{cu} = R_{Fe} \tag{2.54}$$

Additionally, R_{out} is selected in such a way, that

$$R_{out} << r_{cu} \tag{2.55}$$

Since R_{Fe} is a function of l_{air}, so the fulfillment of the condition (2.54) can be achieved through the use of the appropriate air gap. It is called the critical air gap l_{cr}, and the length can be found from (2.51) and (2.54)

$$r_{cu} = (\sigma_h + \sigma_e f)\, m\, f \frac{z^2}{\left(\frac{l_{Fe}}{\mu_{Fe}} + \frac{l_{air}}{\mu_{air}}\right)^2} \tag{2.56}$$

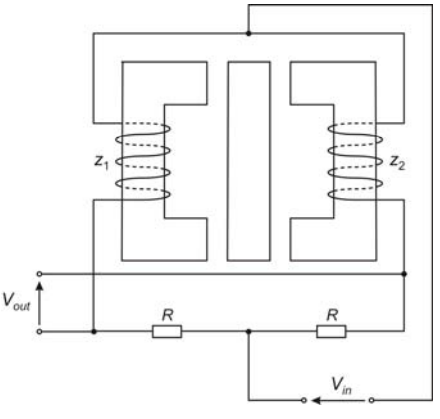

Fig. 2.17 Measuring system arranged as bridge circuit, with inductive sensors in two arms and the common armature

and from it

$$l_{air} = l_{cr} = \mu_{air}\left(\sqrt{\frac{(\sigma_h + \sigma_e f)\, m\, f\, z^2}{r_{cu}}} - \frac{l_{Fe}}{\mu_{Fe}}\right) \tag{2.57}$$

Eq. (2.57) indicates that proper value of the critical length can be obtained through the changes of the number of turns in the coil or mass of the iron circuit.

The measuring system arranged as the bridge, with the inductive sensors in two arms, is shown in Fig. 2.17. The magnetic circuit has the common armature. Before any measurement starts, the armature is in the symmetrical position at the centre. The air gap should be equal to l_{cr}. The armature displacement, one way or the other, results in a push-pull change of magnetic circuit parameters. The balance is distorted and the unbalance voltage V_{out} appears across the output terminals.

The measuring system of four inductive sensors, arranged as the bridge and assigned for the measurement of small angles, is shown in Fig. 2.18. A clockwise turn of the armature makes the impedances Z_1 and Z_4 to decrease, and at the same time to increase the impedances Z_2 and Z_3. The balance is distorted, and the voltage V_{out} appears across the output terminals. The systems of inductive sensors presented and discussed so far are applied to the measurements of the very small displacements within a few millimetres.

Fig. 2.19 shows the most popular variable-inductance sensor, with the movable core, applied for the larger linear-displacement measurements. It is commonly known as the linear variable differential transformer (LVDT).

An LVDT consists of a movable core of magnetic material and three coils, the primary coil and two equal secondaries. The secondary windings are wired in and connected series opposing. Before any measurements start, the core is in a symmetrical position at the centre and the voltages induced in the secondary

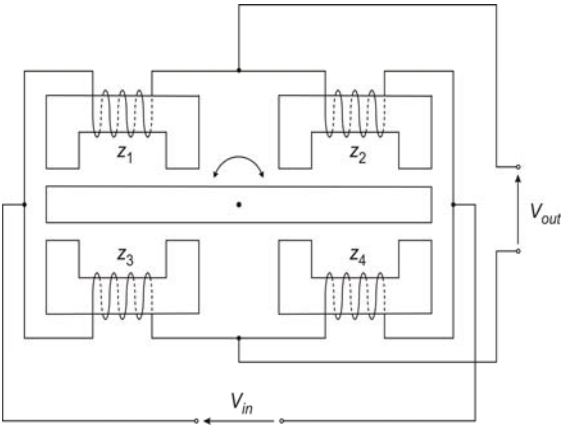

Fig. 2.18 Measuring system containing four inductive sensors and designed for measurement of small angles

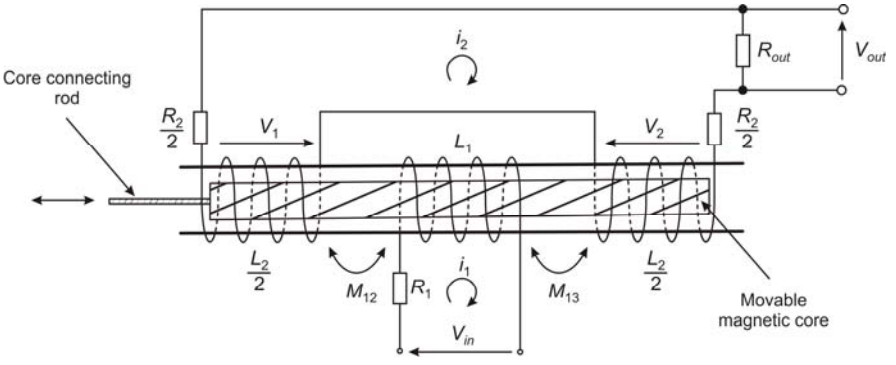

Fig. 2.19 Linear variable differential transformer

coils are equal but out of phase by 180 deg. Since the secondaries are in series opposition, the voltages V_1 and V_2 in the two coils cancel and the output voltage is zero. The displacement of the core leads to an imbalance in mutual inductance between the primary and secondary coils and results in an output voltage development being the difference between V_1 and V_2. Quite often this voltage is applied to the input of the differential amplifier. For the current output we have

$$U_{in} = i_1 (R_1 + j\omega L_1) + j\omega i_2 (M_{13} - M_{12}) \tag{2.58}$$

and

$$i_2 (R + j\omega L_2) + j\omega i_1 (M_{13} - M_{12}) = 0 \tag{2.59}$$

where

$$R = R_2 + R_{out} \tag{2.60}$$

After simple transformation we obtain

$$i_2 = | U_{in} | \frac{\omega (M_{12} - M_{13})}{\sqrt{[RR_1 + \omega^2 (M_{12} - M_{13})^2 - \omega^2 L_1 L_2]^2 + [\omega (R_1 L_2 + R L_1)]^2}} \tag{2.61}$$

2.4 Temperature Sensors

Several temperature sensors are used for temperature measurements, for example liquid sensors, dilatation sensors, bimetal sensors, manometer sensors, semiconductor-based temperature sensors, thermocouples, and resistance devices. However, the emphasis is on thermocouples and resistance devices. They are used very often in all those systems arrangements where temperature is proportional to an electric current or voltage.

A thermoelectric effect, known also as the Seebeck effect, is practically used in thermocouples. A thermocouple is the electric circuit that consist of two dissimilar metals in thermal contact, also called a junction. A thermocouple junction and free, not connected ends of thermocouple wires are maintained at two different temperatures. Generated at the ends of thermocouple a thermal electromotive force (TEF) is proportional to the difference of temperatures. A stability of the reference temperature is a necessary condition of the correct thermocouple operation. To satisfy this condition, lead wires are used for extension of thermocouple wires to the point of the constant temperature. Lead wires should be fabricated from the same pair of metals that are used in the thermocouple. In such a case, no any thermal TEF is generated at the new junctions. If thermocouple wires are produced of different materials, then the new junctions should be maintained at the same temperature.

A thermocouple with lead wires is shown in Fig. 2.20.

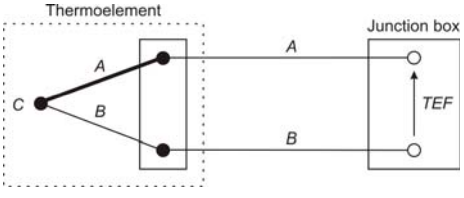

Fig. 2.20 Basic thermocouple circuit A – positive thermocouple wire, B – negative thermocouple wire, C – junction

A large number of materials are suitable for use in thermocouples. Among others, the following pairs are popular combinations of metals to manufacture thermocouples:

- iron vs constantan
- copper vs constantan
- copper vs copper-nickel
- nickel vs chromium-nickel
- nickel vs chromium-constantan
- platinum vs rhodium-platinum.

They are used for a very wide range of temperatures. The thermocouple platinum/rhodium-platinum has some interesting properties. It generates the thermal TEF of 0V between junctions at $0\ ^0$C.

Resistance temperature detectors (RTD) are simply resistive elements of which resistance increases with temperature. In practice, the widely used RTDs are metallic resistors of platinum, nickel and copper.

Semiconductor resistors, also known under the name thermistors, are widely applied for temperature measurements as well. They are fabricated from

semiconductor materials such as oxides of iron, manganese, nickel and lithium and their temperature sensitivity is almost 10 times that of the RTDs. The disadvantage of thermistors is a substantial nonlinearity of characteristics and a significant spread of parameters. It makes the exchange of thermistors difficult in measuring systems, and it is also a reason for a low repeatability of measuring results.

In measurements, the most important are platinum RTDs. Platinum is the superior material for precision thermometry. Platinum RTDs have their mechanical and electrical properties and parameters very stable, and nonlinearity of characteristics is minimal. For this reason, they are used as temperature standards. The usual range of application is up to $1000\,^0\mathrm{C}$, since above this point the resistance of platinum wire changes due to sublimation.

The upper limit of the application of nickel RTDs is determined by the bend of their temperature characteristics, which is around $300\,^0\mathrm{C}$. Copper RTDs are prone to oxidization. They are used mainly in the refrigerating engineering and in the temperature measurements close to ambient temperatures.

Temperature sensors must be protected against mechanical or chemical damages, which may occur during measurements particularly in the industrial environment. For this reason, they are protected through placing them in a thermometer well, which is usually a pipe with a head. Most often, the thermometer wells are fabricated out of cast iron, steel, heat-resisting alloys or ceramic materials, and for obvious reasons they worsen dynamic properties of temperature sensors.

Let us consider the properties of a temperature sensor placed in a single thermometer well, i.e. protected by a single cover. It is shown in Fig. 2.21.

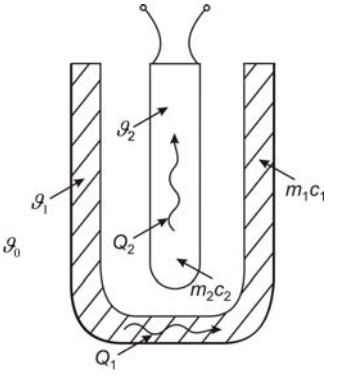

Fig. 2.21 Temperature sensor in a single well

The equation of heat balance is given below

$$dQ = dQ_1 + dQ_2 \tag{2.62}$$

where dQ is the amount of heat that penetrate the device through the well during time dt

$$dQ = (\vartheta_0 - \vartheta_1)k_1 dt \tag{2.63}$$

A part of the total heat dQ is accumulated in the well

$$dQ_1 = m_1 c_1 d\vartheta_1 \tag{2.64}$$

The remaining part is accumulated in the sensor

$$dQ_2 = m_2 c_2 d\vartheta_2 \tag{2.65}$$

Substituting $(2.63)-(2.65)$ into (2.62), we get

$$(\vartheta_0 - \vartheta_1)k_1 dt = m_1 c_1 d\vartheta_1 + m_2 c_2 d\vartheta_2 \tag{2.66}$$

and from that

$$\vartheta_0 = \vartheta_1 + \frac{m_1 c_1}{k_1}\frac{d\vartheta_1}{dt} + \frac{m_2 c_2}{k_1}\frac{d\vartheta_2}{dt} \tag{2.67}$$

At the same time, the amount of heat transferred from the well to the sensor is

$$dQ_2 = (\vartheta_1 - \vartheta_2)k_2 dt \tag{2.68}$$

Substituting (2.65) into (2.68), we get

$$m_2 c_2\, d\vartheta_2 = (\vartheta_1 - \vartheta_2)k_2 dt \tag{2.69}$$

then

$$\vartheta_1 = \vartheta_2 + \frac{m_2 c_2}{k_2}\frac{d\vartheta_2}{dt} \tag{2.70}$$

and

$$\frac{d\vartheta_1}{dt} = \frac{d\vartheta_2}{dt} + \frac{m_2 c_2}{k_2}\frac{d^2\vartheta_2}{dt^2} \tag{2.71}$$

where in $(2.62)-(2.71)$ denotes: $Q(t)-$ the amount of heat penetrating the device, $Q_1(t)-$ the amount of heat warming up the well, $Q_2(t)-$ the amount of heat warming up the sensor, ϑ_0- the measured temperature, ϑ_1- the temperature of the well, ϑ_2- the temperature of the sensor, $m_1 c_1-$ the heat capacity of the well, $m_2 c_2-$ the heat capacity of the sensor, k_1- the heat transfer coefficient of the well, k_2- the heat transfer coefficient from the well to the sensor.

Insertion of (2.70) and (2.71) in (2.67) yields

$$\vartheta_0 = \frac{m_1 c_1}{k_1}\frac{m_2 c_2}{k_2}\frac{d\vartheta_2^2}{dt^2} + \left(\frac{m_1 c_1}{k_1} + \frac{m_2 c_2}{k_2} + \frac{m_2 c_2}{k_1}\right)\frac{d\vartheta_2}{dt} + \vartheta_2 \tag{2.72}$$

Denoting in (2.72)

$$\frac{m_1c_1}{k_1}+\frac{m_2c_2}{k_2}+\frac{m_2c_2}{k_1}=a_1 \quad \text{and} \quad \frac{m_1c_1}{k_1}\frac{m_2c_2}{k_2}=a_2 \tag{2.73}$$

and Laplace transforming both sides of Eq. (2.72), we obtain the transfer function

$$\frac{\vartheta_2(s)}{\vartheta_0(s)}=\frac{1}{a_2s^2+a_1s+1} \qquad a_1,a_2\in\Re \tag{2.74}$$

which has two real and positive poles. They correspond to the time constants T_1 and T_2 of the temperature-measuring device.

$$\frac{\vartheta_2(s)}{\vartheta_0(s)}=\frac{1}{(1+sT_1)(1+sT_2)} \qquad T_1,T_2\in\Re \tag{2.75}$$

Since RTD is a resistive element, the basic circuit for its measurement is a bridge. It can be a full bridge or half a bridge circuit, balanced or unbalanced arrangement. Fig. 2.22 shows an example of the full bridge circuit for the temperature measurements.

Before any measurements, the bridge must be balanced through the appropriate adjustment of the resistor R_2. The resistance of R_{ter} changes together with varying temperature. It causes the bridge to lose the balance. The voltage V_{out} of the unbalance appears across the output terminals. It can be applied to the input of either the current amplifier or the voltage amplifier. In the case of the current amplifier with the adjustable gain, the rated range of the output current is $0-20\text{mA}$. The gain of the amplifier should be adjusted in such a way that the output voltage of the measuring circuits, related to the range of measured temperatures, is within $0-10\text{V}$ range. Nonlinearity of the system characteristics is corrected through the use of the programmable amplifier and appropriate programs. The amplifier cooperates with the output of the bridge.

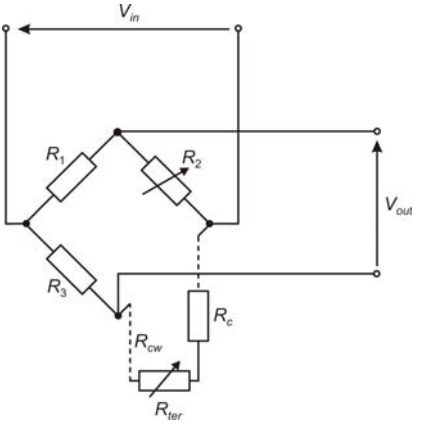

Fig. 2.22 An example of full bridge arranged for temperature measurements

The system works well provided that before a measurement

$$R_{ter} = \frac{R_2 R_3}{R_1} - (R_c + R_{cw}) \tag{2.76}$$

where R_{ter} – thermometer resistance, R_c – compensating resistance, R_{cw} – lead wire resistance and

$$R_c + R_{cw} = R_n = \text{const.} \tag{2.77}$$

where R_n – nominal resistance equals $10\,\Omega$.

If the lead wire resistance R_{cw} changes together with the temperature, $3-$wires asymmetrical balance bridge will be applied. The bridge is shown in Fig. 2.23.

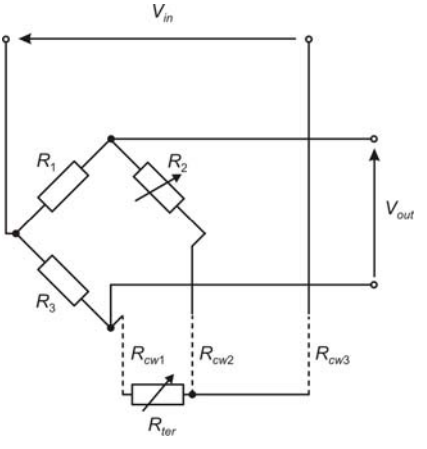

Fig. 2.23 $3-$lead wires asymmetrical balance bridge

For this bridge in balance state we have

$$R_{ter} = \frac{R_2 R_3}{R_1} + \frac{R_3}{R_1} R_{cw2} - R_{cw1} \tag{2.78}$$

For equal resistance of wires

$$R_{cw1} = R_{cw2} = R_{cw} \tag{2.79}$$

Eq. (2.78) becomes

$$R_{ter} = \frac{R_2 R_3}{R_1} + R_{cw}\left(\frac{R_3 - R_1}{R_1}\right) \tag{2.80}$$

The selection of resistors in the bridge should be such that

$$R_3 = R_1 \tag{2.81}$$

then final result given by Eq. (2.80) is as follows

$$R_{ter} = R_2 \tag{2.82}$$

Since a RTD is a resistive element, the basic circuit for its measurements is a bridge. However, during temperature measurements with the use of RTDs, other methods are also used. Quite often the resistance of a RTD can be supplied directly with a constant-current drive. The voltage across R_{ter} is, in such a case, proportional to the resistance value. Nonlinearity of a temperature characteristic is corrected by appropriate software.

2.5 Vibration Sensors

2.5.1 Accelerometer

Sensors intended for measurement of vibration are usually constructed in a form of the damped spring-seismic mass system with a single-degree-of-freedom. Such a seismic sensor model is shown in Fig. 2.24. It can measure either the acceleration or the vibration depending on relations between the seismic mass, damping and the stiffness of the spring.

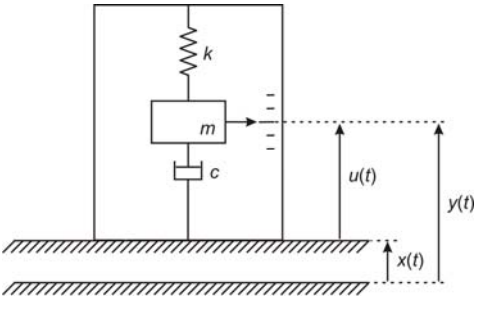

Fig. 2.24 Model of seismic sensor

The model shown in Fig. 2.24 can be described by the equation of motion through the classical second-order differential equation. Using the equation of moments, we can write

$$m\frac{d^2y(t)}{dt^2} + c\frac{du(t)}{dt} + ku(t) = 0 \tag{2.83}$$

where

- $m\frac{d^2y(t)}{dt^2}$ – the moment of inertia
- $c\frac{du(t)}{dt}$ – the moment of dumping
- $ku(t)$ – the moment of elasticity
- $u(t)$ – the relative mass displacement (relative output)

– $y(t)$ – the absolute mass displacement (absolute output)
– m – the seismic mass
– c – dumping coefficient
– k – spring constant.

For the acceleration measurements, the input is the second derivative of absolute displacement $\frac{d^2x(t)}{dt^2}$, and the output is the absolute mass displacement $y(t)$.

From Fig. 2.24, it can be seen that

$$y(t) = u(t) + x(t) \tag{2.84}$$

Substituting (2.84) into (2.83) gives

$$m\frac{d^2u(t)}{dt^2} + c\frac{du(t)}{dt} + ku(t) = -m\frac{d^2x(t)}{dt^2} \tag{2.85}$$

and

$$\frac{1}{\omega_0^2}\frac{d^2u(t)}{dt^2} + \frac{2D}{\omega_0}\frac{du(t)}{dt} + u(t) = -\frac{1}{\omega_0^2}\frac{d^2x(t)}{dt^2} \tag{2.86}$$

where in (2.86)

$$\frac{1}{\omega_0^2} = \frac{m}{k} \quad \text{and} \quad \frac{2D}{\omega_0} = \frac{c}{k} \tag{2.87}$$

ω_0 – undamped natural frequency.

Applying Laplace transform to (2.86), we obtain

$$\frac{1}{\omega_0^2}s^2U(s) + \frac{2D}{\omega_0}sU(s) + U(s) = -\frac{1}{\omega_0^2}s^2X(s) \tag{2.88}$$

The acceleration $s^2X(s)$ is the input signal in s-domain and $U(s)$ is the output. Hence the transfer function $K_{acc}(s)$ of the low-pass accelerometer is

$$K_{acc}(s) = \frac{U(s)}{s^2X(s)} = \frac{-\frac{1}{\omega_0^2}}{\frac{s^2}{\omega_0^2} + \frac{2D}{\omega_0}s + 1} \tag{2.89}$$

The amplitude-frequency characteristic of the transfer function (2.89) is given by Eq. (2.90) and shown in Fig. 2.25

$$\left|K_{acc}(\omega)\right| = \frac{\frac{1}{\omega_0^2}}{\sqrt{\left(1 - \frac{\omega^2}{\omega_0^2}\right)^2 + \left(\frac{2D\omega}{\omega_0}\right)^2}} \tag{2.90}$$

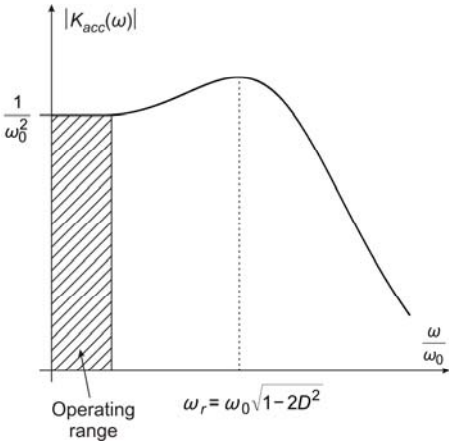

Fig. 2.25 Amplitude-frequency characteristic of accelerometer

The graph presents the amplitude-frequency characteristic of the accelerometer and its operating range. All important harmonics of the measured quantity should be within the range. In practice, the highest frequency value of the operating range should not exceed 33% of the resonance frequency value ω_r.

Examination of Eq. (2.90) indicates that the accelerometer will have a constant gain during the operation when

$$\omega << \omega_0 \tag{2.91}$$

and then

$$\frac{\omega^2}{\omega_0^2} << 1 \quad \text{and} \quad \left(\frac{2D\omega}{\omega_0}\right)^2 << 1 \tag{2.92}$$

Finally

$$\left|K_{acc}(\omega)\right| \to \frac{1}{\omega_0^2} = \text{const.} \tag{2.93}$$

The condition indicated by (2.91) can be achieved through keeping the mass of the accelerometer low and the stiffness of the spring high. For $D = 0.6 - 0.75$, the range of the constant gain has the maximum value.

Regarding the construction, one of possible solutions are accelerometers, the construction of which follows horizontal pendulum kinematics. Their structure includes the mass located on the spring with the strain gauges as the output – Fig. 2.26. The strain gauges located inside the accelerometer are connected into bridges or half a bridges outside the sensor. The accelerometer output signal is the voltage of unbalance in the bridges. Damping is achieved by immersing the system in oil.

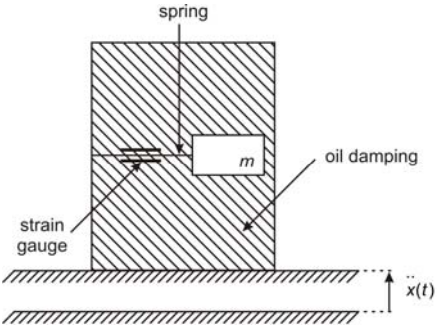

Fig. 2.26 Construction of accelerometer

2.5.2 Vibrometer

The input signal of vibrometers in *s*-domain is the absolute input $X(s)$. Hence the transfer function of the high-pass system is given by

$$K_{wb}(s) = \frac{U(s)}{X(s)} = \frac{-\frac{s^2}{\omega_0^2}}{\frac{s^2}{\omega_0^2} + \frac{2D}{\omega_0}s + 1} \tag{2.94}$$

The amplitude-frequency characteristic of the transfer function (2.94) is given by Eq. (2.95) and shown in Fig. 2.27.

$$\left|K_{wb}(\omega)\right| = \frac{\frac{\omega^2}{\omega_0^2}}{\sqrt{\left(1 - \frac{\omega^2}{\omega_0^2}\right)^2 + \left(\frac{2D\omega}{\omega_0}\right)^2}} \tag{2.95}$$

Examination of Eq. (2.95) indicates that the vibrometer will have a constant gain during the operation when

$$\omega >> \omega_0 \tag{2.96}$$

and then

$$\frac{\omega^2}{\omega_0^2} >> 1 \quad \text{and} \quad \frac{\omega^2}{\omega_0^2} >> \frac{2D\omega}{\omega_0} \tag{2.97}$$

Finally

$$\left|K_{wb}(\omega)\right| \to 1 \tag{2.98}$$

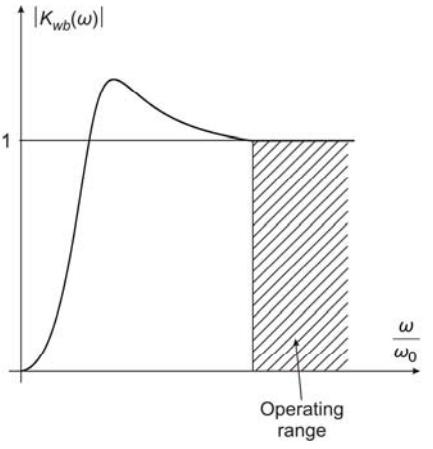

Fig. 2.27 Amplitude-frequency characteristic of vibrometer

It means that the vibrator output is equal to its input. The condition indicated by (2.98) can be achieved through the appropriate design of the vibrator i.e. with soft springs and a relatively large mass, and also with a very small damping.

In practical solutions, a magnetic damping is applied to vibrometers as shown in Fig. 2.28.

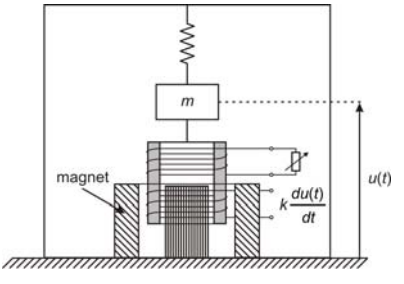

Fig. 2.28 Construction of vibrometer

Two coils are wound up on a bobbin tube, which is mechanically connected to the mass. During vibrations, the coils move into the range of the magnetic field, which exists due to the permanent magnet. A required damping is produced by the coil, which has the adjustable resistor R connected to its output. The voltage induced in the other coil is proportional to $\frac{du(t)}{dt}$. It can be used for velocity measurements or, after integration, it is the output signal of vibrometer.

For both accelerometers and vibrometers, calibration is a process where their amplitude-frequency characteristics are determined. Calibration is carried out using a vibration table with the adjustable amplitude and frequency of vibrations. The high-class frequency generator controls the table. A spiral microscope is the best instrument to determine the amplitude.

2.6 Piezoelectric Sensors

Most piezoelectric sensors are fabricated from quartz crystals SiO_2. There are many advantages of this material. It is very cheap and it exhibits excellent mechanical and electrical properties, high mechanical strength. Its resistivity is high. The influence of temperature variation on the piezoelectric effect in quartz crystals is small.

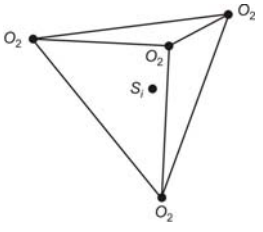

Fig. 2.29 Structure of SiO_2 – silicon dioxide crystal

Fig. 2.29 shows the structure of a silicon dioxide crystal, which is a tetrahedron. A quadrivalent silicon atom Si is inside it and bivalent oxygen atoms O_2 are on the four vertices.

Silicon dioxide crystals interconnect and integrate into monocrystals. These are used as the fundamental material, and plates are cut out from monocrystals. Plates are the basic element for the fabrication of quartz sensors. They are cut out in precise orientation to the crystals axes as shown in Fig. 2.30.

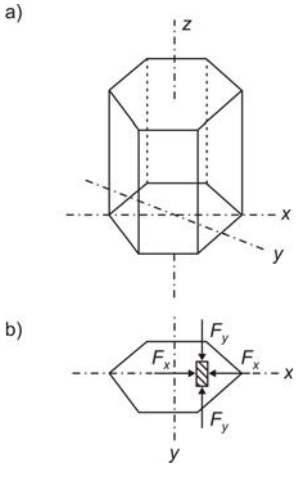

Fig. 2.30 a) Quartz monocrystals SiO_2 b) Preferred axes and orientation of cuts for quartz plates

In the quartz monocrystals, there are three electrical axes, three mechanical ones and one optical axis. The three electrical $x-$axes cross the monocrystals edges and are perpendicular to the optical axis. The three neutral mechanical $y-$axes are perpendicular to the crystal facets. The direction of the optical $z-$axis is such that there is no double refraction for a ray of light along $z-$axis.

A piezoelectric material produces an electric charge when it is subjected to a force or pressure. The main point of piezoelectric effect is that when pressure is applied, or a force causing stretching or compression, the crystal deforms. The deformation produces electric charges on the external surfaces of the crystal and on the metallic electrodes connected to these surfaces. The charges are proportional to the force, which causes the crystal deformation, and they decay when the force is removed

$$Q(t) = k_p F(t) \tag{2.99}$$

where $k_p = 2.3 \cdot 10^{-12} \dfrac{\text{As}}{\text{N}}$ is the piezoelectric constant.

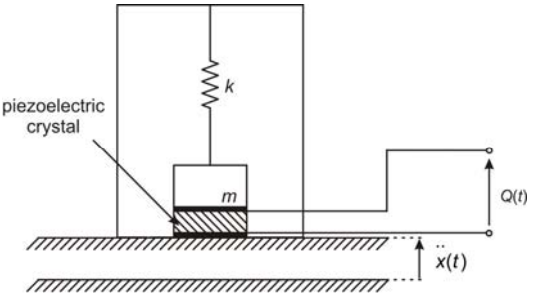

Fig. 2.31 Construction of piezoelectric accelerometer

Fig. 2.31 shows the construction of piezoelectric accelerometer while the circuit diagram of a measuring system with a piezoelectric sensor and a charge amplifier is shown in Fig. 2.32.

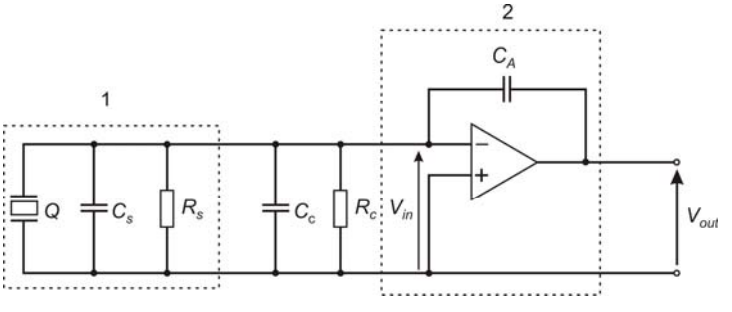

Fig. 2.32 Measuring circuit with piezoelectric sensor $1-$transducer, $2-$charge amplifier, C_s-sensor capacitance, R_s-sensor resistance, $Q-$charge generator, C_c- capacitance of lead wires, R_c-leakance of lead wires

The piezoelectric sensor acts as a charge generator. The charge amplifier consists of a high-gain voltage amplifier with FET at its input for high insulation resistance.

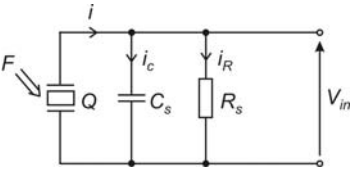

Fig. 2.33 Simplified diagram of a measuring system with a piezoelectric sensor

Fig. 2.33 shows the piezoelectric transducer subjected to a force F that changes sinusoidally. The transducer generates the voltage, which is the input voltage of the amplifier. Let us determine this voltage. For the derivation, C denotes the equivalent capacitance of the transducer and lead wires, and R denotes the equivalent resistance of the same.

Hence we have

$$F(t) = F_m \sin(\omega t) \tag{2.100}$$

and

$$Q(t) = k_p F_m \sin(\omega t) = Q_m \sin(\omega t) \tag{2.101}$$

and also

$$\frac{dQ(t)}{dt} = C\frac{dv(t)}{dt} + \frac{v(t)}{R} \tag{2.102}$$

Therefore

$$\omega Q_m \cos(\omega t) = C\frac{dv(t)}{dt} + \frac{v(t)}{R} \tag{2.103}$$

Laplace transforming Eq. (2.103) results in

$$V(s) = \frac{\omega Q_m}{C}\frac{s}{(s^2+\omega^2)\left(s+\dfrac{1}{RC}\right)} \tag{2.104}$$

The solution of Eq. (2.104) in time-domain is

$$v(t) = \frac{\omega Q_m}{C}\frac{RC}{1+\omega^2R^2C^2}\left[-e^{-\frac{t}{RC}} + \frac{1}{2}(1+j\omega RC)e^{-j\omega t}\frac{1}{2}(1-j\omega RC)e^{j\omega t}\right] \tag{2.105}$$

Eq. (2.105) can be simplified to

$$v(t) = -\frac{\omega Q_m}{C}\frac{RC}{1+\omega^2R^2C^2}e^{-\frac{t}{RC}} + \frac{\omega Q_m}{C}\frac{RC}{1+\omega^2R^2C^2}[\cos(\omega t) + \omega RC\sin(\omega t)] \tag{2.106}$$

After simple trigonometric transformations and taking (2.101) into account, we get finally

$$v(t) = -k_p F_m \frac{\omega R}{1+\omega^2 R^2 C^2} e^{-\frac{t}{RC}} + k_p F_m \frac{\omega R}{\sqrt{1+\omega^2 R^2 C^2}} \sin[\omega t + \frac{\pi}{2} - arctg(\omega RC)] \tag{2.107}$$

Examination of Eq. (2.107) indicates that once the transients have died away the steady-state output signal is the sinusoid with amplitude $k_p F_m \frac{\omega R}{\sqrt{1+\omega^2 R^2 C^2}}$ and phase shift angle $\left(\frac{\pi}{2} - arctg(\omega RC)\right)$.

In the measurements and instrumentation field, the main application of piezoelectric sensors is in construction of piezoelectric quartz accelerometers. In these instruments, the relations between mass m and acceleration a are put to use

$$Q(t) = k_p F(t) = k_p ma \tag{2.108}$$

2.7 Binary-Coded Sensors

The heart of the device is the coded transparent disk. A binary-coded sensor consists of a number of concentric tracks of different diameters. The tracks have a binary coded opaque and transparent pattern. The light source and transducers are perpendicular to the disk. Light generated by photodiodes is detected by photocells, which form a matrix detector. Light passing through a transparent portion is received by a photocell, whereas light blocked by an opaque portion is not received.

Fig. 2.34 presents an example of binary-coded sensors with the binary code and the Gray code. The Gray code is very popular. It shows only a single bit change between adjacent numbers. As a result, the maximum error never exceeds the value of the least significant bit. The advantage of binary-coded sensors is that they are fairly immune from electrical interference. However, they require n tracks for a measurement with the accuracy of $n-$bits, and this is their disadvantage. Most often binary-coded sensors are used for the measurement of angular shaft position.

Incremental angular encoders are another application of this type of sensors. They count segments of a circular graduation and the resulting number denotes a displacement by a specified value. A schematic diagram of the incremental angular encoder is shown in Fig. 2.35.

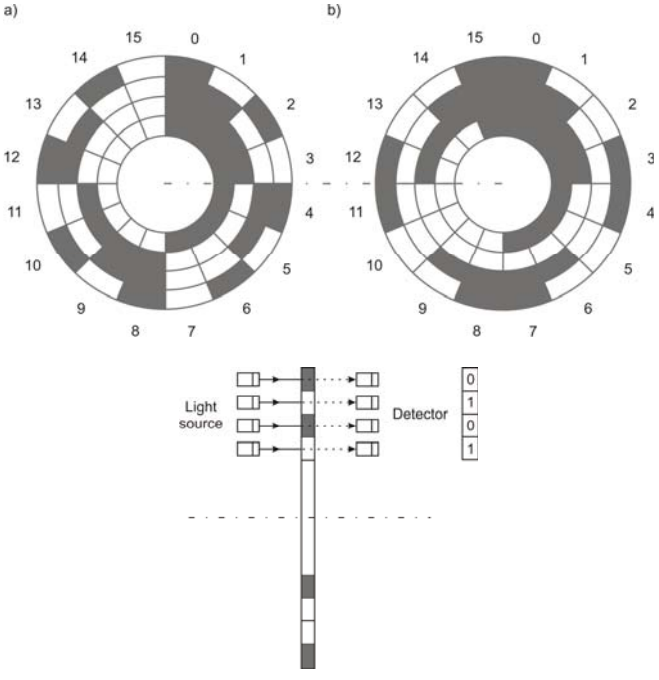

Fig. 2.34 An example of binary-coded sensors with a) binary code and b) Gray code

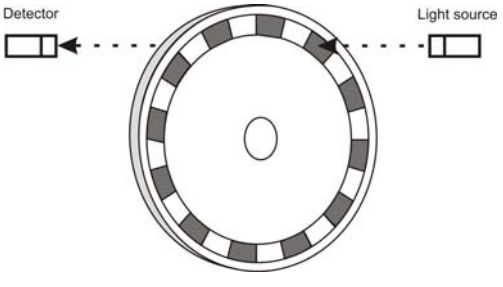

Fig. 2.35 Incremental angular encoder

Opaque and transparent portions are marked along the circumference of the disk, in equal distances from each other. The size of both opaque and transparent portions should be the same. A light source and photodetectors are placed on both sides of the disk, facing one another. Photodetectors measure light intensity of the flux transmitted through the disk.

The system for counting signals is shown in Fig. 2.36. The detector output signal is amplified and modified to TTL standard through the formatting system. The next intermediate stage is the logic gate G with two inputs, the amplified signal from the detector and the standard generator signal. The standard generator

enables or inhibits, logic 1 or 0, the passage of the detector signals through the gate *G*. The counter counts the detector impulses, they are decoded in the decoder and finally displayed in a display unit.

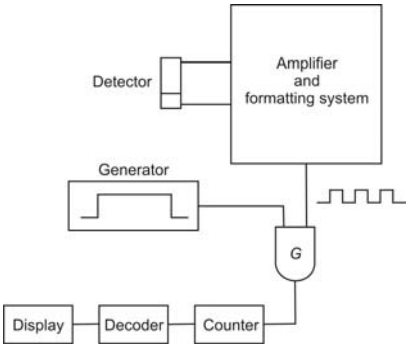

Fig. 2.36 Block diagram of binary-coded transducer

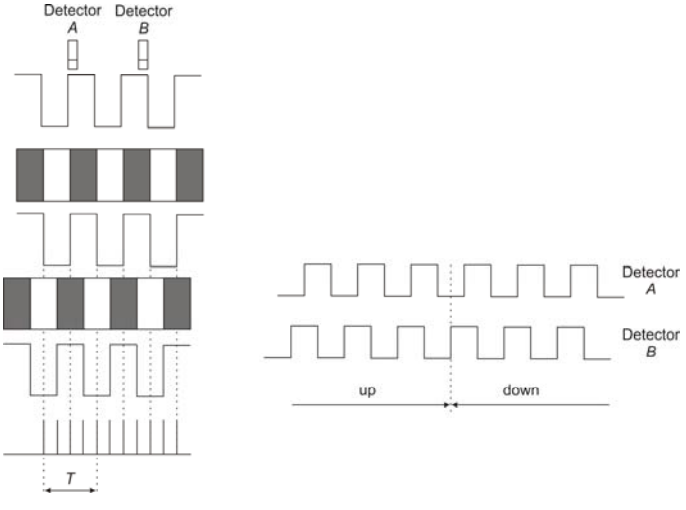

Fig. 2.37 Arrangement of detectors and their output signals

The counted number of impulses is proportional to the rotational speed of the disk. The angular velocity of the disk is given by

$$\omega = \frac{n}{T} = \frac{\theta k}{T} \tag{2.109}$$

where n – rotational speed of the disk, T – time of counting, θ – angle related to one impulse, k – number of impulses.

Two detectors are used to determine the direction of rotations. They are positioned in such a way that the signals generated by them are displaced one towards the other by 90^0. In other words, there is a 90^0-phase angle between them, lagging or leading. Fig. 2.37 shows the arrangement of the detectors and their output signals. When the disk rotates to the right, the counter is counting *up*, while during the rotations to the left, it is counting *down*.

Relations between the detector impulses and the direction of disk rotations:

AB --- 11, 10, 00, 01, 11 --- rotations to the right
AB --- 11, 01, 00, 10, 11 --- rotations to the left

References

[1] Carr, J.: Sensors and Circuits: sensors, transducers, and supporting circuits for electronic instrumentation, measurement, and control. PTR Prentice Hall, Englewood Cliffs (1993)
[2] Nawrocki, W.: Measurement systems and sensors. Artech House, London (2005)
[3] Sinclair, I.R.: Sensors and Transducers. Oxford, Newnes (2001)
[4] Zakrzewski, J.: Czujniki i przetworniki pomiarowe. Gliwice (2004)
[5] Webster, J.G., Pallas Areny, R.: Sensors and Signal Conditioning. Wiley, New York (1991)

Chapter 3
Methods of Noise Reduction

A noise is any unwanted signal mixed in with the desired signal at the output. Noise in a measurement may be from undesired external inputs or generated internally. In some instances, it is very difficult to separate the noise from the measured signal. The error produced by noise can be so significant, in comparison with the measurement signal, that makes the measurement impossible. In particular, it happens in all these cases when the output is the derivative of signals. In case of differentiation of signals, noise is differentiated as well. Due to this it becomes much stronger. Noise reduction by means of filtering, with the use of the weighted mean method, will be discussed in the following subchapters. Especially the Nuttall window, the triangular window and the Kalman filter method will be taken under consideration. In reference to the first case, the analysis of the relations between averaging process and signal distortion will be reviewed as well as the analysis of filtering efficiency in the case of the second-order and third-order objects.

3.1 Weighted Mean Method

Let us consider the object described by the following differential equation

$$\sum_{k=0}^{m} a_k y^{(k)}(t) = u(t) \tag{3.1}$$

where $u(t)$ is the input signal, $y^{(k)}(t)$ is $k-th$ derivative of the output signal, a_k is $k-th$ constant coefficient. The problem to be examined is the determination of the unknown input signal $u(t)$ through the evaluation of the existing signal $y_n(t)$, which is noisy signal.

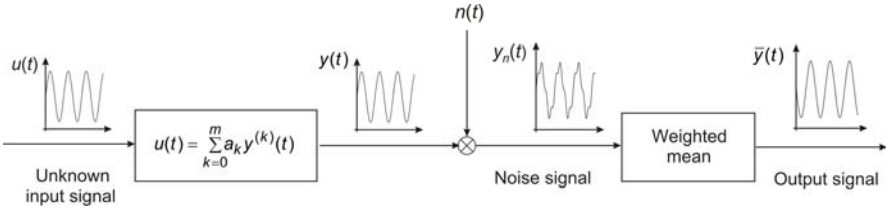

Fig. 3.1 Schematic diagram of weighted mean method

The output signal has two parts. The part desired is due to the unknown input signal $u(t)$ and the undesired part due to all noise inputs

$$y_n(t) = y(t) + n(t) \tag{3.2}$$

In order to determine the input signal $u(t)$, the measured output signal $y(t)$ must be k times differentiated, according to (3.1). The noise output would also be k times differentiated, and as a result the noise would increase significantly. For this reason, the noise should be reduced by filtering before the analogue-to-digital conversion of the signal. Good results of filtering are provided by the weighted mean method that is based on the determination of $\bar{y}(t)$ function

$$\bar{y}(t) = \frac{\int_{t-\delta}^{t+\delta} y_n(\tau)\, g(\tau - t)\, d\tau}{\int_{t-\delta}^{t+\delta} g(\tau - t)\, d\tau} \tag{3.3}$$

where $\bar{y}(t)$ is the weighted mean, $g(\tau - t)$ is the weight function, 2δ is the width of the intervals of averaging.

The properties of averaging depend on the width of the interval 2δ and on the form of the function $g(\tau - t)$. Aiming at filtration, the function $g(\tau - t)$ and its successive derivatives with respect to τ should be equal to zero at the ends of the averaging intervals $(t - \delta)$, $(t + \delta)$

$$g^{(k)}(t - \delta) = g^{(k)}(t + \delta) = 0 \qquad k = 0, 1, 2, \ldots \tag{3.4}$$

and reach the maximum value in the middle of them.

In order to simplify calculations, it is convenient to normalize the denominator of Eq. (3.3). Let

$$d = \int_{t-\delta}^{t+\delta} g(\tau - t)\, d\tau \tag{3.5}$$

then the normalized weighted mean is given as

$$\bar{y}(t) = d^{-1} \int_{t-\delta}^{t+\delta} y_n(\tau)\, g(\tau - t)\, d\tau \tag{3.6}$$

It is easy to check, that the $k-$th derivative of $\bar{y}(t)$ is given by the following equation

$$\bar{y}(t) = (-1)^k\, d^{-1} \int_{t-\delta}^{t+\delta} y_n(\tau)\, g^{(k)}(\tau - t)\; d\tau \tag{3.7}$$

Substituting (3.2) into (3.7) we have

$$\bar{y}_n(t) = (-1)^k\, d^{-1} \int_{t-\delta}^{t+\delta} y(\tau) g^{(k)}(\tau - t)\, d\tau + (-1)^k\, d^{-1} \int_{t-\delta}^{t+\delta} n(\tau) g^{(k)}(\tau - t) d\tau \tag{3.8}$$

in which the differentiation of $n(\tau)$ has been transferred to the function of weight.

Let us estimate the second integral in (3.8)

$$d^{-1}\int_{t-\delta}^{t+\delta} n(\tau)\, g^{(k)}(\tau - t)\, d\tau \;\le\; \sup_{t-\delta\le\tau\le t+\delta} [g^{(k)}(\tau - t)]\, d^{-1}\int_{t-\delta}^{t+\delta} n(\tau)\, d\tau \tag{3.9}$$

Assuming that $n(\tau)$ is the random signal changing quickly its value and the sign with respect to $g^{(k)}(\tau - t)$, we get

$$\int_{t-\delta}^{t+\delta} n(\tau)\, d\tau \approx 0 \tag{3.10}$$

which means noise reduction. The weighted mean of the noise output signal is thus represented by the approximate relation

$$\bar{y}_n \approx (-1)^k d^{-1}\int_{t-\delta}^{t+\delta} y(\tau)\, g^{(k)}(\tau - t)\, d\tau \tag{3.11}$$

3.2 Windows

The requirements with respect to the function $g(\tau - t)$ for which successive derivatives should be equal to zero at the ends of the averaging intervals and reach the maximum value in the middle of them are well fulfilled by Nuttall window and triangular window. The Nuttall window has the form

$$g(\tau - t) = \cos^p\left[\frac{\pi}{2\delta}(\tau - t)\right] \quad p = 1, 2, 3, \ldots \tag{3.12}$$

and is shown in Fig. 3.2.

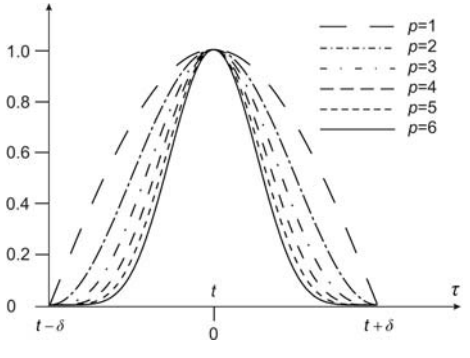

Fig. 3.2 Nuttall windows $g(\tau - t) = \cos^p\left[\frac{\pi}{2\delta}(\tau - t)\right]$ for $p = 1, 2, \ldots, 6$

Table 3.1 presents d values (3.5) of this window.

Table 3.1 Values d of Nuttall window

p	d
1	$\frac{4\delta}{\pi}$
2	δ
3	$\frac{8\delta}{3\pi}$
4	$\frac{3\delta}{4}$
5	$\frac{32\delta}{15\pi}$
6	$\frac{5\delta}{8}$

The triangular window has the form

$$g(\tau - t) = \left[1 - \left|\frac{\tau - t}{\delta}\right|\right]^p \qquad p = 1, 2, 3, \ldots \tag{3.13}$$

and is shown in Fig. 3.3.

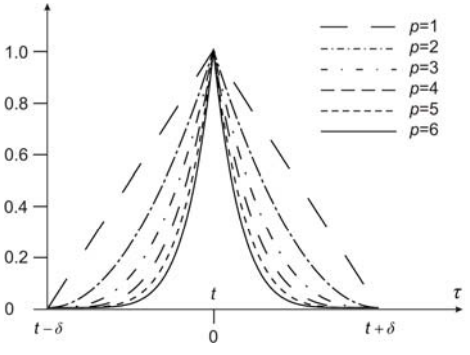

Fig. 3.3 Triangular windows $g(\tau - t) = \left[1 - \left|\frac{\tau - t}{\delta}\right|\right]^p$ for $p = 1, 2, \ldots, 6$

Table 3.2 presents d values (3.5) of this window.

Table 3.2 Values d of triangular window

p	d
1	δ
2	$\frac{2\delta}{3}$

Table 3.2 (*continued*)

3	$\frac{\delta}{2}$
4	$\frac{2\delta}{5}$
5	$\frac{\delta}{3}$
6	$\frac{2\delta}{7}$

3.3 Effect of Averaging Process on Signal Distortion

We will consider the effect of an averaging process on signal distortion, while the windows presented above are applied. Errors generated by the use of windows in the filtering process will also be discussed.

Let us examine the expansion of the continuous signal $f(t)$ into Maclaurin series

$$f(t) = f(0) + \sum_{k=1}^{\infty} \frac{1}{k!} f^{(k)}(0) t^{\mathrm{k}} \tag{3.14}$$

The weighted mean $\bar{f}(t)$ of the signal $f(t)$ is

$$\bar{f}(t) = \frac{1}{\int_{t-\delta}^{t+\delta} g(\tau - t) d\tau} \left[\int_{t-\delta}^{t+\delta} f(0) g(\tau - t) d\tau + \sum_{k=1}^{\infty} f^{(k)}(0) \frac{1}{k!} \int_{t-\delta}^{t+\delta} \tau^{(k)} g(\tau - t) d\tau \right] \tag{3.15}$$

and after simplification it can be written as follows

$$\bar{f}(t) = f(0) + \sum_{k=1}^{\infty} f^{(k)}(0) \frac{1}{k!} \bar{t}^{(k)} \tag{3.16}$$

Using a Nuttall window, we calculate $\bar{f}(t)$ assuming $p = 2$ as an example

$$g(\tau - t) = \cos^2 \left[\frac{\pi}{2\delta} (\tau - t) \right] \tag{3.17}$$

and determine the differences between successive coefficients of the weighted mean $\bar{f}(t)$ (3.16) and the function $f(t)$ (3.14).

For the zero-order derivative ($k = 0$), we get

$$\bar{f}(t) = f(0) \tag{3.18}$$

For the first-order derivative and $k = 1,$ we calculate the integral

$$\frac{f'(0)}{1!} \frac{1}{\delta} \int_{t-\delta}^{t+\delta} \tau \cos^2 \left[\frac{\pi}{2\delta} (\tau - t) \right] d\tau = f \cdot t \tag{3.19}$$

hence the first term of series is not burdened with error.

For the second-order derivative and $k = 2$, we calculate the integral

$$\frac{f''(0)}{2!}\frac{1}{\delta}\int_{t-\delta}^{t+\delta}\tau^2\cos^2\left[\frac{\pi}{2\delta}(\tau - t)\right]d\tau = \frac{f''(0)(\pi^2\delta^2 + 3\pi^2t^2 - 6\delta^2)}{6\pi^2} \tag{3.20}$$

It can be seen that the second term of series is burdened with error

$$\frac{f''(0)(\pi^2\delta^2 - 6\delta^2)}{6\pi^2} \tag{3.21}$$

The successive terms of series (3.16) and the values of error are shown in Table 3.3.

Table 3.3 Successive terms of Maclaurin series and the values of error for Nuttall window, $k = 1, 2, ..., 6$

k		
1	$\bar{f}_k(t)$	$f \cdot t$
	$E_{Nuttall,k}(t,\delta)$	0
2	$\bar{f}_k(t)$	$\frac{f''(0)(\pi^2\delta^2 + 3\pi^2t^2 - 6\delta^2)}{6\pi^2}$
	$E_{Nuttall,k}(t,\delta)$	$\frac{f''(0)(\pi^2\delta^2 - 6\delta^2)}{6\pi^2}$
3	$\bar{f}_k(t)$	$f'''(0)t\frac{\pi^2t^3 + \pi^2\delta^2 - 6\delta^2}{6\pi^2}$
	$E_{Nuttall,k}(t,\delta)$	$f'''(0)t\frac{\pi^2\delta^2 - 6\delta^2}{6\pi^2}$
4	$\bar{f}_k(t)$	$f^{(4)}(0)\frac{\pi^4\delta^4 - 20\pi^2\delta^4 + 10\pi^4t^2\delta^2}{120\pi^4} + f^{(4)}(0)\frac{-60\pi^2t^2\delta^2 + 120\delta^4 + 5\pi^4t^4}{120\pi^4}$
	$E_{Nuttall,k}(t,\delta)$	$f^{(4)}(0)\frac{\pi^4\delta^4 - 20\pi^2\delta^4 + 10\pi^4t^2\delta^2}{120\pi^4} + f^{(4)}(0)\frac{-60\pi^2t^2\delta^2 + 120\delta^4}{120\pi^4}$

Table 3.3 (*continued*)

5	$\bar{f}_k(t)$	$f^{(5)}(0)t\dfrac{3\pi^4\delta^4-60\pi^2\delta^4+360\delta^4}{360\pi^4}$ $+f^{(5)}(0)\dfrac{10\pi^4t^2\delta^2-60\pi^2t^2\delta^2+3\pi^4t^4}{360\pi^4}$
	$E_{Nuttall,k}(t,\delta)$	$f^{(5)}(0)t\dfrac{3\pi^4\delta^4-60\pi^2\delta^4+360\delta^4}{360\pi^4}$ $+f^{(5)}(0)\dfrac{10\pi^4t^2\delta^2-60\pi^2t^2\delta^2}{360\pi^4}$
6	$\bar{f}_k(t)$	$f^{(6)}(0)\dfrac{35\pi^6t^4\delta^2-42\pi^4\delta^6+840\pi^2\delta^6}{5040\pi^6}$ $+f^{(6)}(0)\dfrac{-210\pi^4t^4\delta^2+2520\pi^2t^2\delta^4+21\pi^6t^2\delta^4}{5040\pi^6}$ $+f^{(6)}(0)\dfrac{-420\pi^4t^2\delta^4+7\pi^6t^6-5040\delta^6+\pi^6\delta^6}{5040\pi^6}$
	$E_{Nuttall,k}(t,\delta)$	$f^{(6)}(0)\dfrac{35\pi^6t^4\delta^2-42\pi^4\delta^6+840\pi^2\delta^6}{5040\pi^6}$ $+f^{(6)}(0)\dfrac{-210\pi^4t^4\delta^2+2520\pi^2t^2\delta^4+21\pi^6t^2\delta^4}{5040\pi^6}$ $+f^{(6)}(0)\dfrac{-420\pi^4t^2\delta^4-5040\delta^6+\pi^6\delta^6}{5040\pi^6}$

Fig. 3.4 shows the total error $E_{Nuttall}(t,\delta)$ equal to the sum of error components $E_{Nuttall,k}(t,\delta)$ listed in Table 3.3.

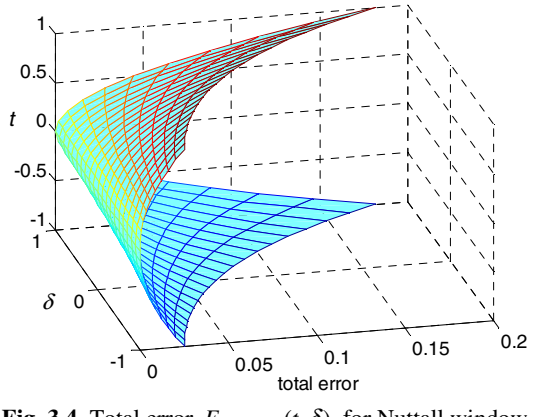

Fig. 3.4 Total error $E_{Nuttall}(t,\delta)$ for Nuttall window

The value of error and the successive terms of series (3.16) for the triangular window for

$$g(\tau - t) = \left[1 - \left|\frac{\tau - t}{\delta}\right|\right]^2 \tag{3.22}$$

and $k = 1, 2, ..., 6$ are shown in Table 3.4.

Table 3.4 Successive terms of Maclaurin series for triangular window, $k = 1, 2, ..., 6$

k		
1	$\bar{f}_k(t)$	$f \cdot t$
	$E_{Triangular,k}(t,\delta)$	0
2	$\bar{f}_k(t)$	$\frac{1}{20} f^{(2)}(0)(\delta^2 + 10t^2)$
	$E_{Triangular,k}(t,\delta)$	$\frac{1}{20} f^{(2)}(0)\delta^2$
3	$\bar{f}_k(t)$	$\frac{1}{60} f^{(3)}(0)t(3\delta^2 + 10t^2)$
	$E_{Triangular,k}(t,\delta)$	$\frac{1}{20} f^{(3)}(0)t\delta^2$
4	$\bar{f}_k(t)$	$\frac{1}{840} f^{(4)}(0)(35t^4 + 21t^2\delta^2 + \delta^4)$
	$E_{Triangular,k}(t,\delta)$	$\frac{1}{840} f^{(4)}(0)(21t^2\delta^2 + \delta^4)$
5	$\bar{f}_k(t)$	$\frac{1}{840} f^{(5)}(0)t(7t^4 + 7t^2\delta^2 + \delta^4)$
	$E_{Triangular,k}(t,\delta)$	$\frac{1}{840} f^{(5)}(0)t(7t^2\delta^2 + \delta^4)$
6	$\bar{f}_k(t)$	$\frac{1}{60480} f^{(6)}(0)(84t^6 + 126t^4\delta^2 + 36t^2\delta^4 + \delta^6)$
	$E_{Triangular,k}(t,\delta)$	$\frac{1}{60480} f^{(6)}(0)(126t^4\delta^2 + 36t^2\delta^4 + \delta^6)$

Fig. 3.5 shows the total error $E_{Triangular}(t,\delta)$ equal to the sum of error components $E_{Triangular,k}(t,\delta)$ listed in Table 3.4.

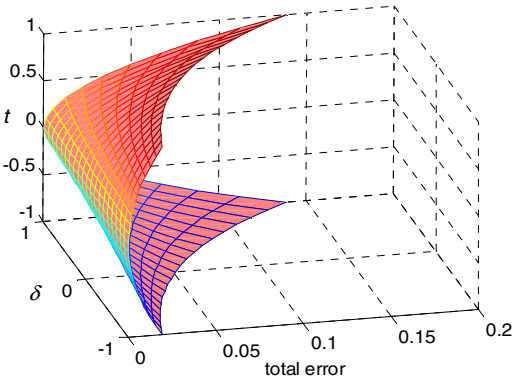

Fig. 3.5 Total error $E_{Triangular}(t,\delta)$ for triangular window

Fig. 3.6 shows a comparison of errors from Fig. 3.4 and Fig. 3.5.

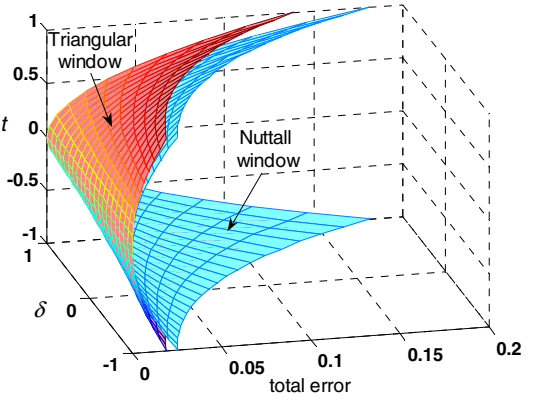

Fig. 3.6 Comparison of sum of errors for Nuttall window and triangular window

The maximum values of errors for the Nuttall window Fig. 3.4 and the triangular window Fig. 3.5 are as follows

$$\sup_{t,\,\delta} E_{Nuttall}(t,\delta) = 0.181 \quad \text{for} \quad t \in [-1,\,1],\ \delta \in [-1,\,1]$$

$$\sup_{t,\,\delta} E_{Triangular}(t,\delta) = 0.138 \quad \text{for} \quad t \in [-1,\,1],\ \delta \in [-1,\,1]$$

The comparison of these results indicates that errors generated during the averaging process are smaller in the case of the triangular windows than in the Nuttall windows.

3.4 Efficiency Analysis of Noise Reduction by Means of Filtering

The noise reduction efficiency when using filtering will be analysed on the examples of the second and third-order objects as well as Nuttall and triangular windows.

Let the second order object be given in the following form

$$a_2 y''(t) + a_1 y'(t) + a_0 y(t) = u(t) \tag{3.23}$$

where $u(t)$ is the input and $y(t)$ the output signal as mentioned in (3.1). Let the output signal be the sinusoid

$$y(t) = Y \sin(\omega t + \nu) \tag{3.24}$$

After substitution of (3.24) into (3.23) and simple calculations of derivatives, we get

$$u(t) = a_0 Y \sin(\omega t + \nu) + a_1 Y \omega \cos(\omega t + \nu) - a_2 Y \omega^2 \sin(\omega t + \nu) \tag{3.25}$$

If the output signal $y(t)$ is mixed with noise, we will use the weighted mean $\overline{u}(t)$ instead of $u(t)$. Replacing $y(t)$ by $u(t)$ and substituting (3.23) into (3.6), we get

$$\overline{u}(t) = (-1)^k d^{-1} \int_{t-\delta}^{t+\delta} [(a_2 y''(\tau) + a_1 y'(\tau) + a_0 y(\tau)) g^{(k)}(\tau - t)] d\tau \tag{3.26}$$

It can be shown as the three separate integrals

$$\begin{aligned} \overline{u} = {} & (-1)^2 d^{-1} \int_{t-\delta}^{t+\delta} a_2 y''(\tau) g^{(k)}(\tau - t) d\tau \\ & + (-1)^1 d^{-1} \int_{t-\delta}^{t+\delta} a_1 y'(\tau) g^{(k)}(\tau - t) d\tau \\ & + (-1)^0 d^{-1} \int_{t-\delta}^{t+\delta} a_0 y(\tau) g^{(k)}(\tau - t) d\tau \end{aligned} \tag{3.27}$$

Transferring the respective derivatives from $y''(\tau)$ and $y'(\tau)$ to the weight functions $g''(\tau - t)$ and $g'(\tau - t)$, we have

$$\begin{aligned} \overline{u} = {} & d^{-1} \int_{t-\delta}^{t+\delta} a_2 y(\tau) g''(\tau - t) d\tau - d^{-1} \int_{t-\delta}^{t+\delta} a_1 y(\tau) g'(\tau - t) d\tau \\ & + d^{-1} \int_{t-\delta}^{t+\delta} a_0 y(\tau) g(\tau - t) d\tau \end{aligned} \tag{3.28}$$

Applying Nuttall window (3.12), let us recalculate (3.28). The successive derivatives for Nuttall window

$$g(\tau - t) = \cos^p\left[\frac{\pi}{2\delta}(\tau - t)\right] \tag{3.29}$$

are as follows

$$g'(\tau - t) = \frac{-p\pi}{2\delta}\sin\left[\frac{\pi}{2\delta}(\tau - t)\right]\cos^{p-1}\left[\frac{\pi}{2\delta}(\tau - t)\right] \tag{3.30}$$

$$g''(\tau - t) = \frac{-p\pi^2}{4\delta^2}\cos^{p-2}\left[\frac{\pi}{2\delta}(\tau - t)\right]\left[-p + p\cos^2\left[\frac{\pi}{2\delta}(\tau - t)\right] + 1\right] \tag{3.31}$$

Substituting (3.29) – (3.31) into (3.28) gives

$$\begin{aligned}
&\bar{u}_{Nuttall}(t) = \\
&-d^{-1}\int_{t-\delta}^{t+\delta} a_2 y(\tau)\frac{p\pi^2}{4\delta^2}\cos^{p-2}\left[\frac{\pi}{2\delta}(\tau - t)\right]\left[-p + p\cos^2\left[\frac{\pi}{2\delta}(\tau - t)\right] + 1\right]d\tau \\
&+d^{-1}\int_{t-\delta}^{t+\delta} a_1 y(\tau)\frac{p\pi}{2\delta}\sin\left[\frac{\pi}{2\delta}(\tau - t)\right]\cos^{p-1}\left[\frac{\pi}{2\delta}(\tau - t)\right]d\tau \\
&+d^{-1}\int_{t-\delta}^{t+\delta} a_0 y(\tau)\cos^p\left[\frac{\pi}{2\delta}(\tau - t)\right]d\tau
\end{aligned} \tag{3.32}$$

Considering (3.24), we have

$$\begin{aligned}
&\bar{u}_{Nuttall}(t) = \\
&\quad d^{-1}\int_{t-\delta}^{t+\delta} a_2 Y\sin(\omega\tau + \nu)\frac{p\pi^2}{4\delta^2}\cos^{p-2}\left[\frac{\pi}{2\delta}(\tau - t)\right]p\,d\tau \\
&\quad -d^{-1}\int_{t-\delta}^{t+\delta} a_2 Y\sin(\omega\tau + \nu)\frac{p\pi^2}{4\delta^2}\cos^{p-2}\left[\frac{\pi}{2\delta}(\tau - t)\right] \\
&\quad \cdot\left[p\cos^2\left[\frac{\pi}{2\delta}(\tau - t)\right] + 1\right]d\tau \\
&\quad +d^{-1}\int_{t-\delta}^{t+\delta} a_1 Y\sin(\omega\tau + \nu)\frac{p\pi}{2\delta}\sin\left[\frac{\pi}{2\delta}(\tau - t)\right]\cos^{p-1}\left[\frac{\pi}{2\delta}(\tau - t)\right]d\tau \\
&\quad +d^{-1}\int_{t-\delta}^{t+\delta} a_0 Y\sin(\omega\tau + \nu)\cos^p\left[\frac{\pi}{2\delta}(\tau - t)\right]d\tau
\end{aligned} \tag{3.33}$$

Calculating the integrals in (3.33), for $p = 2$ and $d^{-1} = \frac{1}{\delta}$ as an example, we get finally

$$\begin{aligned}
\bar{u}_{Nuttall}(t) = &[a_0 Y\sin(\omega t + \nu) + a_1 Y\omega\cos(\omega t + \nu) - a_2 Y\omega^2\sin(\omega t + \nu)] \\
&\cdot\frac{\pi^2}{\pi^2 - (\omega\delta)^2}\frac{\sin(\omega\delta)}{\omega\delta}
\end{aligned} \tag{3.34}$$

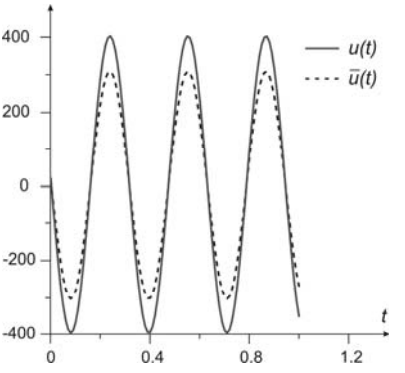

Fig. 3.7 Filtering efficiency of Nuttall window for second-order object

It is easy to see the difference between signals (3.25) and (3.34). For $\omega\delta = \text{const}$ magnitude of the signal $\bar{u}(t)$ is multiplied by the constant coefficient, in comparison with the signal $u(t)$. The value of this coefficient decreases to zero, if $\omega\delta$ tends to infinity. Fig. 3.7 presents the diagrams of signal $u(t)$ and $\bar{u}(t)$.

Let us repeat the similar analysis for the triangular window. Substituting the weight function (3.13), its respective derivatives $g'(\tau - t)$, $g''(\tau - t)$ and the output signal (3.24) into (3.28), we get

$$\begin{aligned}
\bar{u}_{Triangular}(t) = {} & d^{-1}\int_{t-\delta}^{t} a_2 Y \sin(\omega\tau + \nu)\frac{(p^2-p)}{\delta^2}\left(1+\frac{\tau-t}{\delta}\right)^{p-2} d\tau \\
& + d^{-1}\int_{t}^{t+\delta} a_2 Y \sin(\omega\tau + \nu)\frac{(p^2-p)}{\delta^2}\left(1-\frac{\tau-t}{\delta}\right)^{p-2} d\tau \\
& - d^{-1}\int_{t-\delta}^{t} a_1 Y \sin(\omega\tau + \nu)\frac{p}{\delta}\left(1+\frac{\tau-t}{\delta}\right)^{p-1} d\tau \\
& + d^{-1}\int_{t}^{t+\delta} a_1 Y \sin(\omega\tau + \nu)\frac{p}{\delta}\left(1-\frac{\tau-t}{\delta}\right)^{p-1} d\tau \\
& + d^{-1}\int_{t-\delta}^{t} a_0 Y \sin(\omega\tau + \nu)\left(1+\frac{\tau-t}{\delta}\right)^{p} d\tau \\
& + d^{-1}\int_{t}^{t+\delta} a_0 Y \sin(\omega\tau + \nu)\left(1-\frac{\tau-t}{\delta}\right)^{p} d\tau
\end{aligned} \tag{3.35}$$

Substituting $p = 2$ and $d^{-1} = \dfrac{3}{2\delta}$ into (3.35) and integrating, we finally have

$$\begin{aligned}
\bar{u}_{Triangular}(t) = {} & \frac{6}{(\omega\delta)^3}[a_0 Y \sin(\omega t + \nu)[\omega\delta - \sin(\omega\delta)] + a_1 Y\omega\cos(\omega t + \nu) \\
& \cdot [\omega\delta - \sin(\omega\delta)] + a_2 Y\omega^2 \sin(\omega t + \nu)\sin(\omega\delta)]
\end{aligned} \tag{3.36}$$

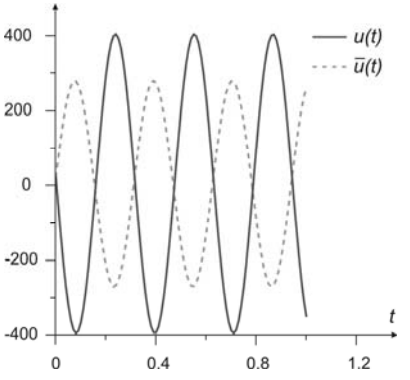

Fig. 3.8 Filtering efficiency of triangular window for second-order object

Fig. 3.8. presents the diagrams of the signals $u(t)$ and $\bar{u}(t)$. In this case, the difference between the signals refers both to the magnitude and phase displacement. The latter one equals π rad.

Let us check now the efficiency of filtering in the case of Nuttall and triangular window application to a third-order object. Let this object be given in the following form

$$a_3 y'''(t) + a_2 y''(t) + a_1 y'(t) + a_0 y(t) = u(t) \tag{3.37}$$

The output signal is the same as given by Eq (3.24). Substituting (3.24) into (3.37) yields

$$\begin{aligned} u(t) = &-a_3 Y\omega^3 \cos(\omega t + \nu) - a_2 Y\omega^2 \sin(\omega t + \nu) \\ &+ a_1 Y\omega \cos(\omega t + \nu) + a_0 Y \sin(\omega t + \nu) \end{aligned} \tag{3.38}$$

For the third-order object (3.37) the weighted mean $\bar{u}(t)$ has the following form

$$\begin{aligned} \bar{u}(t) = &-d^{-1} \int_{t-\delta}^{t+\delta} a_3 y(\tau) g'''(\tau - t)\, d\tau \\ &+ d^{-1} \int_{t-\delta}^{t+\delta} a_2 y(\tau) g''(\tau - t)\, d\tau \\ &- d^{-1} \int_{t-\delta}^{t+\delta} a_1 y(\tau) g'(\tau - t)\, d\tau \\ &+ d^{-1} \int_{t-\delta}^{t+\delta} a_0 y(\tau) g(\tau - t)\, d\tau \end{aligned} \tag{3.39}$$

Taking Nuttall window (3.12) under consideration and calculating the respective derivatives of (3.39), we get

$$
\begin{aligned}
&\bar{u}_{Nuttall}(t) = \\
&d^{-1}\int_{t-\delta}^{t+\delta} a_3 Y \sin(\omega\tau+\nu)\frac{p\pi^3}{8\delta^3}\sin\left[\frac{\pi}{2\delta}(\tau-t)\right]\cos\left[\frac{\pi}{2\delta}(\tau-t)\right]^{p-3} p^2 d\tau \\
&-d^{-1}\int_{t-\delta}^{t+\delta} a_3 Y \sin(\omega\tau+\nu)\frac{p\pi^3}{8\delta^3}\sin\left[\frac{\pi}{2\delta}(\tau-t)\right]\cos\left[\frac{\pi}{2\delta}(\tau-t)\right]^{p-3} \\
&\cdot p^2\cos\left[\frac{\pi}{2\delta}(\tau-t)\right]^2 d\tau \\
&-d^{-1}\int_{t-\delta}^{t+\delta} a_3 Y \sin(\omega\tau+\nu)\frac{p\pi^3}{8\delta^3}\sin\left[\frac{\pi}{2\delta}(\tau-t)\right]\cos\left[\frac{\pi}{2\delta}(\tau-t)\right]^{p-3}[3p-2]d\tau \\
&-d^{-1}\int_{t-\delta}^{t+\delta} a_2 Y \sin(\omega\tau+\nu)\frac{p\pi^2}{4\delta^2}\cos^{p-2}\left[\frac{\pi}{2\delta}(\tau-t)\right] \\
&\cdot\left[-p+p\cos^2\left[\frac{\pi}{2\delta}(\tau-t)\right]+1\right]d\tau \\
&+d^{-1}\int_{t-\delta}^{t+\delta} a_1 Y \sin(\omega\tau+\nu)\frac{p\pi}{2\delta}\sin\left[\frac{\pi}{2\delta}(\tau-t)\right]\cos^{p-1}\left[\frac{\pi}{2\delta}(\tau-t)\right]d\tau \\
&+d^{-1}\int_{t-\delta}^{t+\delta} a_0 Y \sin(\omega\tau+\nu)\cos^{p}\left[\frac{\pi}{2\delta}(\tau-t)\right]d\tau
\end{aligned}
\tag{3.40}
$$

From the expression (3.40), for $p=3$ and $d^{-1}=\dfrac{3\pi}{8\delta}$, we get

$$
\begin{aligned}
\bar{u}_{Nuttall}(t) = [&a_0 Y\sin(\omega t+\nu)+a_1 Y\omega\cos(\omega t+\nu)-a_2 Y\omega^2\sin(\omega t+\nu) \\
&-a_3 Y\omega^3\cos(\omega t+\nu)]\frac{9\pi^4\cos(\omega\delta)}{\omega^2\delta^2(16\omega^2\delta^2-40\pi^2)-9\pi^4}
\end{aligned}
\tag{3.41}
$$

The results of the calculations are shown in a form of diagrams in Fig 3.9.

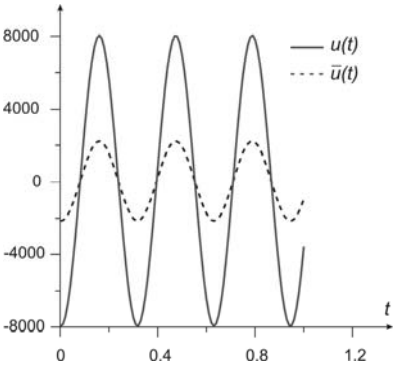

Fig. 3.9 Filtering efficiency of Nuttall window for third-order object

When comparing the latter and the former results, it is evident that the ratio of the voltage $u(t)$ and $\bar{u}(t)$ magnitudes depends on the order of object.

For triangular window (3.13) and the third-order object, Eq (3.39) takes the form

$$\begin{aligned}
&\bar{u}_{Triangular}(t) = \\
&-d^{-1}\int_{t-\delta}^{t} a_3 Y\sin(\omega\tau+\nu)\frac{(p^3-3p^2+2p)}{\delta^3}\left(1+\frac{\tau-t}{\delta}\right)^{p-3}d\tau \\
&-d^{-1}\int_{t}^{t+\delta} a_3 Y\sin(\omega\tau+\nu)\frac{(-p^3+3p^2-2p)}{\delta^3}\left(1-\frac{\tau-t}{\delta}\right)^{p-3}d\tau \\
&+d^{-1}\int_{t-\delta}^{t} a_2 Y\sin(\omega\tau+\nu)\frac{(p^2-p)}{\delta^2}\left(1+\frac{\tau-t}{\delta}\right)^{p-2}d\tau \\
&+d^{-1}\int_{t}^{t+\delta} a_2 Y\sin(\omega\tau+\nu)\frac{(p^2-p)}{\delta^2}\left(1-\frac{\tau-t}{\delta}\right)^{p-2}d\tau \\
&-d^{-1}\int_{t-\delta}^{t} a_1 Y\sin(\omega\tau+\nu)\frac{p}{\delta}\left(1+\frac{\tau-t}{\delta}\right)^{p-1}d\tau \\
&+d^{-1}\int_{t}^{t+\delta} a_1 Y\sin(\omega\tau+\nu)\frac{p}{\delta}\left(1-\frac{\tau-t}{\delta}\right)^{p-1}d\tau \\
&+d^{-1}\int_{t-\delta}^{t} a_0 Y\sin(\omega\tau+\nu)\left(1+\frac{\tau-t}{\delta}\right)^{p}d\tau \\
&+d^{-1}\int_{t}^{t+\delta} a_0 Y\sin(\omega\tau+\nu)\left(1-\frac{\tau-t}{\delta}\right)^{p}d\tau
\end{aligned} \tag{3.42}$$

Substituting $p=3$ and $d^{-1}=\frac{2}{\delta}$, we finally get

$$\begin{aligned}
&\bar{u}_{Triangular}(t) = \frac{24}{\omega^4\delta^4}\left[\frac{(\omega\delta)^2}{2}+\cos(\omega\delta)-1\right]a_0 Y\sin(\omega t+\nu) \\
&-\frac{24}{\omega^4\delta^4}(a_1 Y\omega\sin(\omega t+\nu)[\omega\delta-\sin(\omega\delta)] \\
&+a_2 Y\omega^2\sin(\omega t+\nu)[\cos(\omega\delta)-1]+a_3 Y\omega^3\cos(\omega t+\nu)[\cos(\omega\delta)-1])
\end{aligned} \tag{3.43}$$

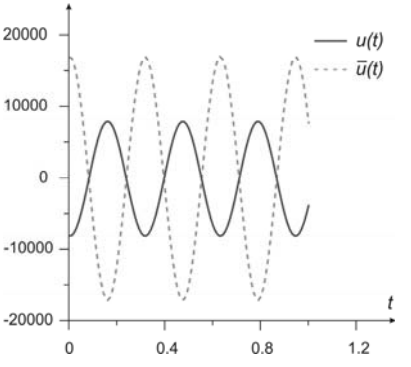

Fig. 3.10 Filtering efficiency of triangular window for third-order object

The phase displacement of the signals $u(t)$ and $\overline{u}(t)$ is π rad, likewise in the second-order object. The ratio of voltage magnitudes depends on the order of object, in a similar way like in the case of Nuttall window.

3.5 Kalman Filter

So far the reduction of noise by filtering, with the application of the weighted mean methods, has been discussed. Kalman filter method is another quite popular way, often used in practice, to achieve this aim. It is applied to a linear discrete dynamic object. For such a object, the recurrent algorithm of minimum variance estimator of the state vector is being developed. This aim is achieved through the use of the output of dynamic object given by the discrete state equations

$$\begin{aligned} \mathbf{x}(k+1) &= \mathbf{A}(k)\,\mathbf{x}(k) + \mathbf{B}(k)\,\mathbf{u}(k) + \mathbf{w}(k) \\ \mathbf{y}(k) &= \mathbf{C}(k)\,\mathbf{x}(k) + \mathbf{D}(k)\,\mathbf{u}(k) + \mathbf{v}(k) \qquad k = 0,1,2,\ldots \end{aligned} \tag{3.44}$$

For Kalman filter, it is assumed that both the measurement and the conversion process inside the object are burdened with an error described by the standardized normal distribution. Fig. 3.11 shows the block diagram of the object represented by Eq. (3.44)

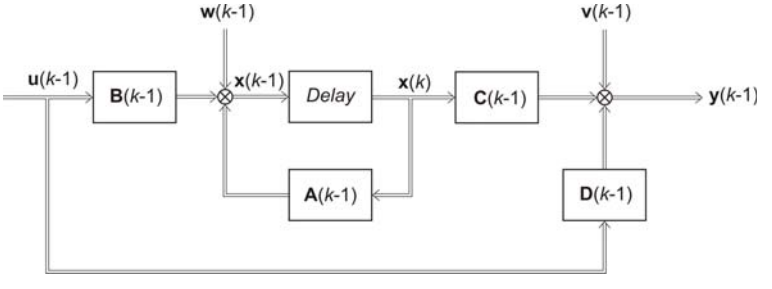

Fig. 3.11 Block diagram of discrete dynamic object

$\mathbf{u}(k)$ – vector of input signal of m dimension, $\mathbf{x}(k)$ and $\mathbf{x}(k+1)$ – state vectors of n dimension at time k and $k+1$, $\mathbf{y}(k)$ – vector of output signal of p dimension, $\mathbf{w}(k)$ – vector of object noise of n dimension, $\mathbf{v}(k)$ – measurement noise vector of p dimension, $\mathbf{A}(k)_{nxn}$ – state matrix, $\mathbf{B}(k)_{nxm}$ – input matrix, $\mathbf{C}(k)_{pxn}$ – output matrix, $\mathbf{D}(k)_{pxm}$ – direct transmission matrix

The following assumptions are introduced for the synthesis of Kalman filter:

1. The deterministic component of the input signal $\mathbf{u}(k)$ equals zero
2. In case of control lack, the state variable oscillates around zero

$$E[\mathbf{x}(k)] = 0 \tag{3.45}$$

3. Noises $\mathbf{w}(k)$ and $\mathbf{v}(k)$ both have properties of discrete white noise. It means they are not correlated, their expected value is zero and their covariance is constant

$$E[\mathbf{w}(k)\mathbf{w}^T(k)]=\begin{cases}\mathbf{R}(k) & \text{if} \quad i=k\\ 0 & \text{if} \quad i\neq k\end{cases} \tag{3.46}$$

$$E[\mathbf{v}(k)\mathbf{v}^T(k)]=\begin{cases}\mathbf{Q}(k) & \text{if} \quad i=k\\ 0 & \text{if} \quad i\neq k\end{cases} \tag{3.47}$$

where $\mathbf{R}(k)$ and $\mathbf{Q}(k)$ are the covariance matrices of noise.

4. The state errors and the measurement errors are not correlated

$$E[\mathbf{v}(k)\mathbf{w}^T(k)]=0 \tag{3.48}$$

5. The estimation errors do not depend on measurements

$$E[(\mathbf{x}(k)-\hat{\mathbf{x}}(k))\,\mathbf{v}^T(k)]=0 \tag{3.49}$$

It means that the vector $\hat{\mathbf{x}}(k)$ depends on the observation vector at random. The relation holds until $k-1$ step.

6. The matrix $\mathbf{D}(k)=0$

Such assumption enables to modify the state equation (3.44) to the following form

$$\begin{aligned}\mathbf{x}(k+1)&=\mathbf{A}(k)\mathbf{x}(k)+\mathbf{B}(k)\mathbf{u}(k)\\ \mathbf{y}(k)&=\mathbf{C}(k)\mathbf{x}(k)+\mathbf{v}(k)\end{aligned} \tag{3.50}$$

The block diagram related to the above equation is shown in Fig. 3.12.

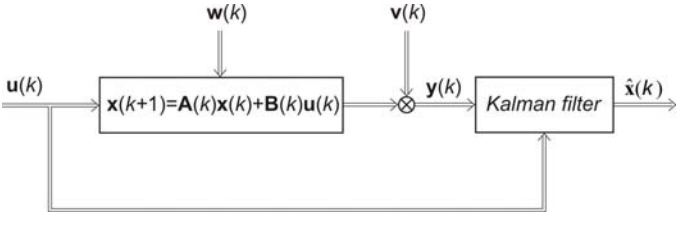

Fig. 3.12 Schematic diagram of Kalman filtering

The idea of Kalman filter is based on the assumption that the linear state estimator $\hat{\mathbf{x}}(k-1,k-1)$ and the covariance $\mathbf{P}(k-1,k-1)$ can be obtained through $k-1$ observations of the object output at the discrete instant $k-1$. The next step is prediction of the values of both the estimator $\hat{\mathbf{x}}(k,k-1)$ and the covariance

$\mathbf{P}(k,k-1)$, the latter tied in with the former, at the time instant k. If there is a difference between the obtained results and those predicted during the previous step, a correction must be made to the prediction for the instant $k+1$. The correction is carried out at the time instant k.

Kalman filter equations are based on these assumptions. They are divided into two categories (i) and (ii), described below in details.

(i) Equations of time updating

On the basis of the estimation at the instant $k-1$, the prediction is done at the discrete instant k. The time updating equations enable the prediction. The following algorithm complete the task:

1. Project the state ahead

$$\hat{\mathbf{x}}(k,k-1)=\mathbf{A}(k)\hat{\mathbf{x}}(k-1,k-1)+\mathbf{B}(k)\mathbf{u}(k-1) \tag{3.51}$$

where $\hat{\mathbf{x}}(k-1,k-1)$ and $\hat{\mathbf{x}}(k,k-1)$ are the corresponding estimations of the state vector before and after the measurement

2. Project the error covariance ahead

$$\mathbf{P}(k,k-1)=\mathbf{A}(k)\mathbf{P}(k-1,k-1)\mathbf{A}^T(k)+\mathbf{R}(k) \tag{3.52}$$

where

$$\mathbf{P}(k-1,k-1)=\mathbf{E}[\,\mathbf{e}(k-1,k-1)\mathbf{e}^T(k-1,k-1)] \tag{3.53}$$

is the covariance matrix of the a priori error vector

$$\mathbf{e}(k-1,k-1)=\mathbf{x}(k-1)-\hat{\mathbf{x}}(k-1,k-1) \tag{3.54}$$

and

$$\mathbf{P}(k,k-1)=\mathbf{E}[e(k,k-1)\mathbf{e}^T(k,k-1)] \tag{3.55}$$

where

$$\mathbf{e}(k,k-1)=\mathbf{x}(k)-\hat{\mathbf{x}}(k,k-1) \tag{3.56}$$

is the covariance matrix of the a posteriori error vector.

The difference between the real value of the state vector and its estimation is presented by the vectors (3.54) and (3.56). This difference is a good measure of the error of the state vector assessment.

(ii) Equations of measurements updating

On the basis of the actual observation data, the prediction is a corrected by the measurement updating equations. The algorithm of procedure is as follows:

1. Compute the Kalman gain

$$\mathbf{K}(k)=\mathbf{P}(k,k-1)\mathbf{C}^T(k)[\mathbf{Q}(k)+\mathbf{C}(k)\mathbf{P}(k,k-1)\mathbf{C}^T(k)]^{-1} \tag{3.57}$$

2. Update the estimate with measurement $\mathbf{y}(k)$

$$\hat{\mathbf{x}}(k) = \hat{\mathbf{x}}(k, k-1) + \mathbf{K}(k, k)[\mathbf{y}(k) - \mathbf{C}(k)\,\hat{\mathbf{x}}(k, k-1)] \tag{3.58}$$

3. Update the error covariance

$$\mathbf{P}(k) = [\mathbf{I} - \mathbf{K}(k, k)\mathbf{C}(k)]\mathbf{P}(k, k-1) \tag{3.59}$$

The algorithm presenting the whole action and operation of Kalman filter, following the equations (3.51) to (3.59), is shown in Fig. 3.13.

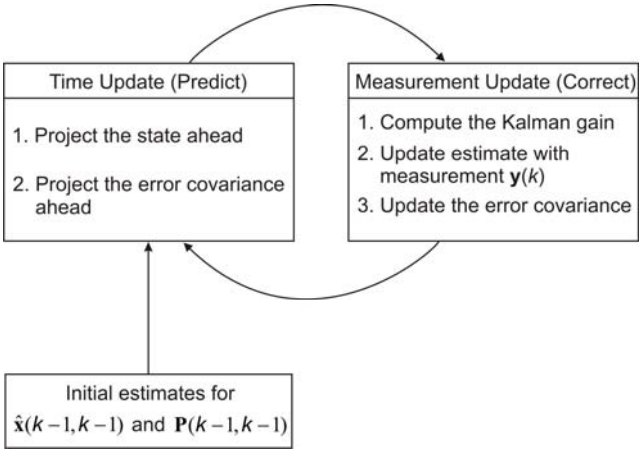

Fig. 3.13 Algorithm of Kalman filter operation

During the operation of Kalman filter, the equations of updating time and of measurements work in cycles, in the successive instants k between one action and another. It makes possible to estimate the process of $\hat{\mathbf{x}}(k)$ according to the minimum of mean-square error.

For numerical calculations, the initial parameters should be taken under considerations. Either there is some preliminary information about the process or the assumption must be made about zero initial conditions. The latter case refers to the state vector estimate. Additionally, the covariance matrix $\mathbf{P}(k-1, k-1)$ should have large value elements. If too small values of matrix elements are assumed, it will results in the gain matrix $\mathbf{K}(k)$ being small in the consecutive steps, and the estimates $\hat{\mathbf{x}}(k)$ will be close to the initial values. Further consequence of such an approach is that the optimum solution will only be obtained after a significant increase of the number of iteration steps. On the other hand, if too large values of the covariance matrix elements are assumed, the estimate $\hat{\mathbf{x}}(k)$ will change quickly in reference to its initial value. It will be seen in the form of a significant overshoot during the initial step of estimation.

References

[1] Friedland, B.: On the properties of reduced-order Kalman filters. IEEE Trans. Autom. Control. 34, 321–324 (1989)
[2] Grewal, M.S.: Kalman filtering. Theory & Practice. Prentice Halls, Englewood Cliffs (1993)
[3] Kordylewski, W., Wach, J.: Usrednione rozniczkowanie zakloconych sygnalow pomiarowych. Pomiary Automatyka Kontrola 6, 123–124 (1988)
[4] Layer, E.: Modelling of Simplified Dynamical Systems. Springer, Heidelberg (2002)
[5] Nuttall, A.H.: Some windows with very good sidelobe behaviour. IEEE Trans. on Acoustic, Speech and Signal Processing 29(1) (1981)
[6] Orlowski, M.: Odtwarzanie usrednionych sygnalow wejsciowych na podstawie zaszumianych sygnalow wyjsciowych. Phd Thesis, Politechnika Szczecinska, Szczecin (1992)
[7] Zuchowski, A.: O pomiarach charakterystyk dynamicznych obiektow w warunkach zaklocen. Pomiary Automatyka Kontrola 11, 273–275 (1991)
[8] Zuchowski, A.: Dynamic measurement of the curve y(x) defined by parametric relations x(t), y(t) under random disturbances. Metrology and Measurement Systems XI(2) (2005)
[9] Zuchowski, A.: Przyczynek do metod filtracji sygnalow. Pomiary Automatyka Kontrola 2, 7–8 (2006)
[10] Zuchowski, A., Grzywacz, B.: Koncepcja ukladu dla filtracji zakłocen z jednoczesnym wiernym odtwarzaniem duzych skokowych zmian sygnalu. Pomiary Automatyka Kontrola 53, 3–4 (2007)

Chapter 4
Model Development

Selected methods of development of various time-invariant models are presented in the chapter.

Using algebraic polynomials, approximation methods are reviewed. The polynomials of Lagrange, Tchebychev, Legendre and Hermite are studied in detail. These methods are used quite often provided that the number of data points is not too large. That is because the order of the polynomial is equal to the number of data. Too large number of data results in an equally high number of the polynomial order.

When the approximations of functions having irregular waveforms are considered, it is convenient to apply the cubic splines approximation method. It is based on splitting the given interval into a collection of subintervals, followed by the approximation of the data at each subinterval by means of the cubic order polynomial. The method is described in the following parts of the chapter in detail.

Another method, which is discussed in the chapter, makes possible a derivation of a relatively low degree polynomial, which will pass "near" the measured data points instead of passing through them. It is the least squares approximation method for which the error being a sum of squares of the differences between the values of the approximation line and the measured data is at minimum. Approximation by means of power series, with the use of Maclaurin series, is presented in the next part of this chapter. This method is particularly useful in the case of models in dynamic state because Maclaurin series describes a function near the origin. There is also an additional advantage of the method. Coefficients of the series can be transformed directly into state equations coefficients or coefficients of Laplace transfer functions. These two forms are applied most often in modelling various objects of electrical and control engineering. There are a couple of other methods, which are discussed in the following parts of the chapter. The standard nets method, which allows for an easy the determination of the order of a modelled object, and the optimization method based on Levenberg-Marguardt algorithm with LabVIEW program application, are presented. Finally,

the black-box identification for discrete models in the form of ARX with the MATLAB program application and the Monte Carlo method are also considered.

4.1 Lagrange Polynomials

Let us consider the polynomial $L(x)$

$$L(x) = a_0 L_0(x) + a_1 L_1(x) + a_2 L_2(x) + ... + a_{n-1} L_{n-1}(x) \tag{4.1}$$

If in (4.1)

$$L_k(x) = x^k, \quad k = 0,1,...,n-1 \tag{4.2}$$

then $L(x)$ is called the Lagrange interpolating polynomial. Polynomial $L(x)$ at each measuring point x_k fulfills the condition

$$L_k(x_k) = f(x_k) \tag{4.3}$$

where $f(x_k)$ presents measuring data in x_k. Six graphs of the first consecutive polynomials $L_k(x)$ are shown in Fig. 4.1

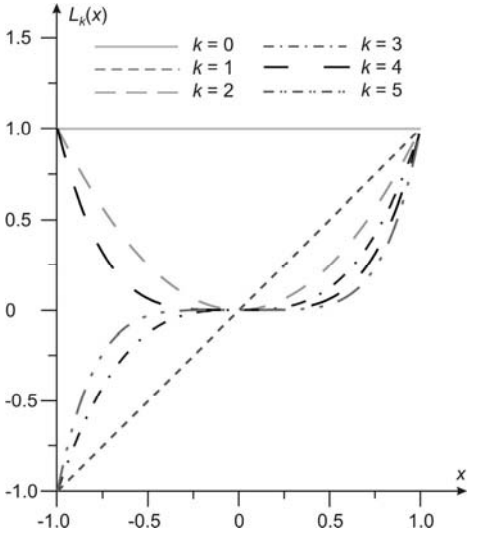

Fig. 4.1 The first six Lagrange polynomials

In order to determine unknown coefficients $a_0, a_1, ..., a_{n-1}$ of the polynomial $L(x)$, let us substitute Eq. (4.3) into Eq. (4.1). Thus we have the following system of *n* linear equations

$$\begin{aligned}
&a_0 + a_1 x_0 + a_2 x_0^2 + \ldots + a_{n-1} x_0^{n-1} = f(x_0) \\
&a_0 + a_1 x_1 + a_2 x_1^2 + \ldots + a_{n-1} x_1^{n-1} = f(x_1) \\
&\cdot \quad \cdot \quad \cdot \quad \cdot \quad \cdot \\
&\cdot \quad \cdot \quad \cdot \quad \cdot \quad \cdot \\
&a_0 + a_1 x_{n-1} + a_2 x_{n-1}^2 + \ldots + a_{n-1} x_{n-1}^{n-1} = f(x_{n-1})
\end{aligned} \tag{4.4}$$

The system of equations (4.4) can be presented in matrix form

$$\begin{bmatrix}
1 & x_0 & x_0^2 & . & . & x_0^{n-1} \\
1 & x_1 & x_1^2 & . & . & x_1^{n-1} \\
. & . & . & . & . & . \\
. & . & . & . & . & . \\
1 & x_{n-1} & x_{n-1}^2 & . & . & x_{n-1}^{n-1}
\end{bmatrix}
\begin{bmatrix} a_0 \\ a_1 \\ . \\ . \\ a_{n-1} \end{bmatrix} =
\begin{bmatrix} f(x_0) \\ f(x_1) \\ . \\ . \\ f(x_{n-1}) \end{bmatrix} \tag{4.5}$$

where the vector of coefficients $\boldsymbol{a}$ presents the solution.

The matrix on the left is known as a Vandermonde matrix. It has the non-zero determinant, which indicates that the system (4.5) has a solution for $\boldsymbol{a}$, and the solution is unique.

Let us consider the cardinal function

$$C_k(x) = \frac{\prod_{k=0}^{n-1} (x - x_k)}{(x - x_k) \dfrac{d}{dx} \prod_{k=0}^{n-1} (x - x_k) \Big|_{x = x_k}} \tag{4.6}$$

that has the following properties

$$C_k(x_i) = \begin{cases} 1 & \text{if} \quad k = i \\ 0 & \text{if} \quad k \neq i \end{cases} \tag{4.7}$$

After a simple transformation, relation (4.6) can be presented in the form

$$C_k(x) = \prod_{\substack{i=0 \\ i \neq k}}^{n-1} \frac{(x - x_i)}{(x_k - x_i)} \tag{4.8}$$

also occurring in other polynomials e.g. Tchebychev, Legendre, Hermite etc.

The interpolation polynomial $L(x)$, presented by means of (4.8), takes the form

$$L(x) = \sum_{k=0}^{n-1} f(x_k) C_k(x) \tag{4.9}$$

Fig. 4.2 shows the components of polynomial (4.9) for five exemplary measuring points $(x, f(x)) = (1,3;\ 2,5;\ 3,2;\ 4,4;\ 5,3)$ in the interval [1, 5].

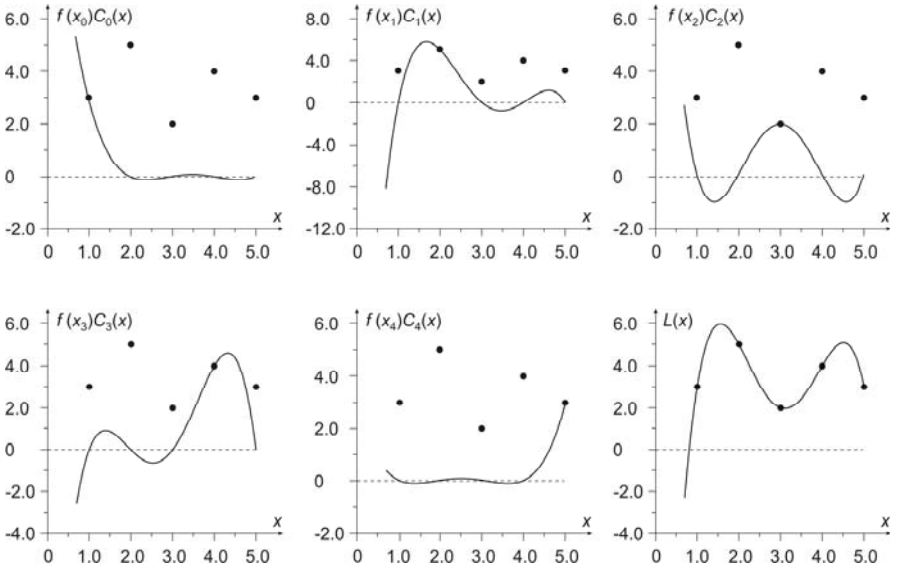

Fig. 4.2 Exemplary components of polynomial (4.9)

4.2 Tchebychev Polynomials

In (4.1), let us replace the polynomial $L(x)$ by Tchebychev polynomials $T(x)$

$$T(x) = a_0 T_0(x) + a_1 T_1(x) + a_2 T_2(x) + \ldots + a_{n-1} T_{n-1}(x) \tag{4.10}$$

For each measuring point x_k we have

$$T_k(x_k) = f(x_k), \quad k = 0,1,\ldots,n-1 \tag{4.11}$$

The individual polynomials occurring in (4.10) can be determined with the use of the recurrence formula

$$\begin{aligned} T_{k+1}(x) &= 2xT_k(x) - T_{k-1}(x) \\ T_0(x) &= 1 \\ T_1(x) &= x \end{aligned} \tag{4.12}$$

Some of the initial Tchebychev polynomials are given by

$$\begin{aligned} T_0(x) &= 1 \\ T_1(x) &= x \\ T_2(x) &= 2x^2 - 1 \\ T_3(x) &= 4x^3 - 3x \\ T_4(x) &= 8x^4 - 8x^2 + 1 \\ T_5(x) &= 16x^5 - 20x^3 + 5x \end{aligned} \tag{4.13}$$

and are shown in Fig. 4.3

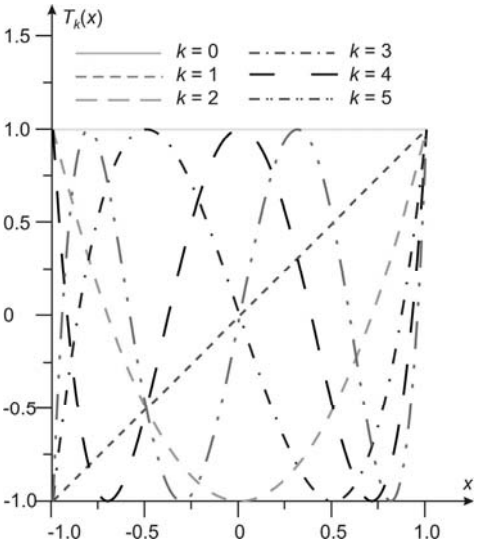

Fig. 4.3 The first six Tchebychev polynomials

After substituting Eq. (4.11) into Eq. (4.10) and taking (4.12) into account, the system of equations (4.14) can be obtained, where the vector of coefficients ***a*** presents the solution

$$\begin{bmatrix} 1 & x_0 & 2x_0^2 - 1 & . & . & 2xT_{n-1}(x_0) - T_{n-2}(x_0) \\ 1 & x_1 & 2x_1^2 - 1 & . & . & 2xT_{n-1}(x_1) - T_{n-2}(x_1) \\ . & . & . & . & . & . \\ . & . & . & . & . & . \\ 1 & x_{n-1} & 2x_{n-1}^2 - 1 & . & . & 2xT_{n-1}(x_{n-1}) - T_{n-2}(x_{n-1}) \end{bmatrix} \begin{bmatrix} a_0 \\ a_1 \\ . \\ . \\ a_{n-1} \end{bmatrix} = \begin{bmatrix} f(x_0) \\ f(x_1) \\ . \\ . \\ f(x_{n-1}) \end{bmatrix} \tag{4.14}$$

The interpolation points x_k, which determine the zeros of $T_k(x)$ in the interval $[-1, 1]$, form a triangular matrix called the experiment plan according to the zeros of the Tchebychev polynomials. For the polynomials (4.13), we have

$$\begin{aligned}
&k=1, \quad x_1=0\\
&k=2, \quad x_2=-\sqrt{2}/2, \quad \sqrt{2}/2\\
&k=3, \quad x_3=-\sqrt{3}/2, \quad 0, \quad \sqrt{3}/2\\
&k=4, \quad x_4=-1/2\sqrt{2-\sqrt{2}}, \quad -1/2\sqrt{2+\sqrt{2}}, \quad 1/2\sqrt{2-\sqrt{2}}, \quad 1/2\sqrt{2+\sqrt{2}}\\
&k=5, \quad x_5=-1/4\sqrt{10+2\sqrt{5}}, \quad -1/4\sqrt{10-2\sqrt{5}},\\
&\qquad 0, \quad 1/4\sqrt{10-2\sqrt{5}}, \quad 1/4\sqrt{10+2\sqrt{5}}
\end{aligned} \tag{4.15}$$

The cardinal functions in the zeros of the Tchebychev polynomials have the form

$$C_k(x)=\frac{T_k(x)}{(x-x_k)\dfrac{d}{dx}T_k(x)\Big|_{x=x_k}} \tag{4.16}$$

for which the polynomial (4.10) can be presented as

$$T(x)=\sum_{k=0}^{n-1} f(x_k)\,C_k(x) \tag{4.17}$$

Fig. 4.4 shows the components of (4.17) for five exemplary measuring points, which are determined by zeros of the fifth order polynomial and by measuring data $f(x_k)$ equal 3, 5, 2, 4, 3, respectively

$$[x, f(x)]=\Big[-1/4\sqrt{10+2\sqrt{5}},\ 3;\quad -1/4\sqrt{10-2\sqrt{5}},\ 5;\quad 0,\ 2;$$
$$1/4\sqrt{10-2\sqrt{5}},\ 4;\quad 1/4\sqrt{10+2\sqrt{5}},\ 3\Big]$$

On the grounds of well-known properties of orthogonal functions, it is the advantage to use orthogonal polynomials in many cases of approximation. In the interval $[-1, 1]$, Tchebychev polynomials are orthogonal with the weight function w (x)

$$w_{[-1,\,1]}(x)=\frac{1}{\sqrt{1-x^2}} \tag{4.18}$$

for which we have

$$\int_{-1}^{1} T_n(x)T_m(x)\frac{dx}{\sqrt{1-x^2}}=\begin{cases}0 & \text{if} \quad n\neq m\\ \pi & \text{if} \quad n=m=0\\ \pi/2 & \text{if} \quad n=m\neq 0\end{cases} \tag{4.19}$$

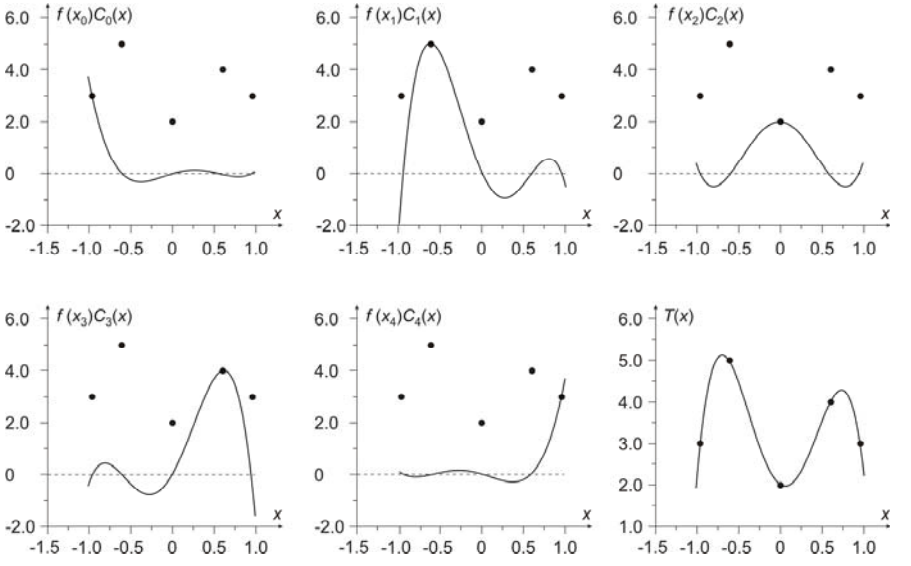

Fig. 4.4 Exemplary components of polynomial (4.18)

Assuming the interpolation points belong to the interval $[a, b]$, they can be transformed into the interval $[-1, 1]$ using the following formula

$$x' = \frac{2x - a - b}{b - a} \tag{4.20}$$

From (4.20), it can be easily noticed that shifted Tchebychev polynomials in the interval $[0, b]$ are presented by

$$\begin{aligned} T_{k+1}(x) &= 2(-1 + 2x/b)T_k(-1 + 2x/b) - T_{k-1}(-1 + 2x/b) \\ T_0(x) &= 1 \\ T_1(x) &= -1 + 2x/b \end{aligned} \tag{4.21}$$

The polynomials (4.21) are orthogonal with the weight function

$$w_{[0,\,b]}(x) = \frac{1}{\sqrt{bx - x^2}} \tag{4.22}$$

A few shifted Tchebychev polynomials in the interval $[0, b]$ are given by

$$\begin{aligned} T_0(x) &= 1 \\ T_1(x) &= -1 + 2x/b \\ T_2(x) &= 1 - 8x/b + 8x^2/b^2 \\ T_3(x) &= -1 + 18x/b - 48x^2/b^2 + 32x^3/b^3 \\ T_4(x) &= 1 - 32x/b + 160x^2/b^2 - 256x^3/b^3 + 128x^4/b^4 \\ T_5(x) &= -1 + 50x/b - 400x^2/b^2 + 1120\,x^3/b^3 - 1280x^4/b^4 + 512\,x^5/b^5 \end{aligned} \tag{4.23}$$

4.3 Legendre Polynomials

In (4.1), let us replace the polynomial $L(x)$ by Legendre polynomials $P(x)$

$$P(x) = a_0 P_0(x) + a_1 P_1(x) + a_2 P_2(x) + ... + a_{n-1} P_{n-1}(x) \tag{4.24}$$

For each measuring point x_k, we have

$$P_k(x_k) = f(x_k), \quad k = 1, 2, ..., n-1 \tag{4.25}$$

The individual polynomials occurring in (4.24) can be determined with the use of the recurrence formula

$$\begin{aligned} P_{k+1}(x) &= \frac{1}{k+1}[(2k+1)\,x\,P_k(x) - k\,P_{k-1}(x)] \\ P_0(x) &= 1 \\ P_1(x) &= x \end{aligned} \tag{4.26}$$

Legendre polynomials are orthogonal in the interval $[-1, 1]$ with the weight function $w(x) = 1$, and fulfill the following condition

$$\int_{-1}^{1} P_n(x) P_m(x) = \begin{cases} 0 & \text{if} \quad n \neq m \\ \dfrac{2}{2n+1} & \text{if} \quad n = m \end{cases} \tag{4.27}$$

Some of the initial Legendre polynomials in the interval $[-1, 1]$ are given by

$$\begin{aligned} P_0(x) &= 1 \\ P_1(x) &= x \\ P_2(x) &= \frac{1}{2}(3x^2 - 1) \\ P_3(x) &= \frac{1}{2}(5x^3 - 3x) \\ P_4(x) &= \frac{1}{8}(35x^4 - 30x^2 + 3) \\ P_5(x) &= \frac{1}{8}(63x^5 - 70x^3 + 15x) \end{aligned} \tag{4.28}$$

and are shown in Fig. 4.5.

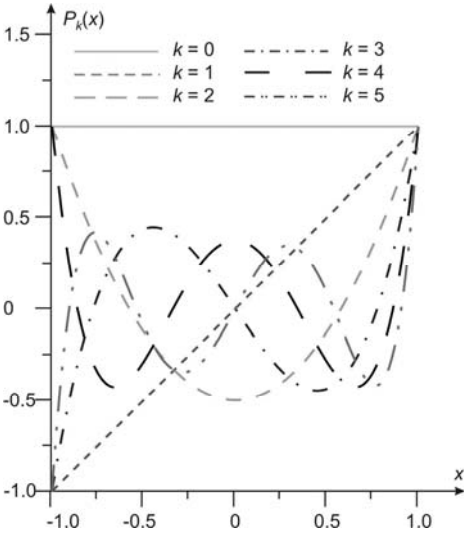

Fig. 4.5 The first five Legendre polynomials

After substituting Eq. (4.25) into Eq. (4.24) and taking (4.26) into account, the system of equations (4.29) can be obtained, where the vector of coefficients $\boldsymbol{a}$ presents the solution

$$\begin{bmatrix} 1 & x_0 & \frac{1}{2}(3x_0^2-1) & \cdots & \frac{1}{n-1}\left[(2n-2)x_0 P_{n-2}(x_0)-(n-2)P_{n-3}(x_0)\right] \\ 1 & x_1 & \frac{1}{2}(3x_1^2-1) & \cdots & \frac{1}{n-1}\left[(2n-2)x_1 P_{n-2}(x_1)-(n-2)P_{n-3}(x_1)\right] \\ \cdot & \cdot & \cdot & \cdots & \cdot \\ \cdot & \cdot & \cdot & \cdots & \cdot \\ 1 & x_{n-1} & \frac{1}{2}(3x_{n-1}^2-1) & \cdots & \frac{1}{n-1}\left[(2n-2)x_0 P_{n-2}(x_{n-1})-(n-2)P_{n-3}(x_{n-1})\right] \end{bmatrix} \cdot \begin{bmatrix} a_0 \\ a_1 \\ \cdot \\ \cdot \\ a_{n-1} \end{bmatrix} = \begin{bmatrix} f(x_0) \\ f(x_1) \\ \cdot \\ \cdot \\ f(x_{n-1}) \end{bmatrix} \tag{4.29}$$

Replacing $T_k(x)$ in (4.16) by $P_k(x)$ of (4.26) gives

$$C_k(x) = \frac{P_k(x)}{(x-x_k)\dfrac{d}{dx}P_k(x)\Big|_{x=x_k}} \tag{4.30}$$

for which the polynomial (4.24) can be presented as

$$P(x) = \sum_{k=0}^{n-1} f(x_k)\, C_k(x) \tag{4.31}$$

Fig. 4.6 shows the components of polynomial (4.31) for five exemplary measuring points, which are determined by zeros of the fifth order polynomial and by measuring data $f(x_k)$ equal 3, 5, 2, 4 and 3, respectively

$$[x, f(x)] = \left[-\sqrt{\frac{5}{9} + \frac{2}{63}\sqrt{70}},\ 3;\quad -\sqrt{\frac{5}{9} - \frac{2}{63}\sqrt{70}},\ 5;\quad 0,\ 2;\right.$$
$$\left. \sqrt{\frac{5}{9} - \frac{2}{63}\sqrt{70}},\ 4;\quad \sqrt{\frac{5}{9} + \frac{2}{63}\sqrt{70}},\ 3 \right]$$

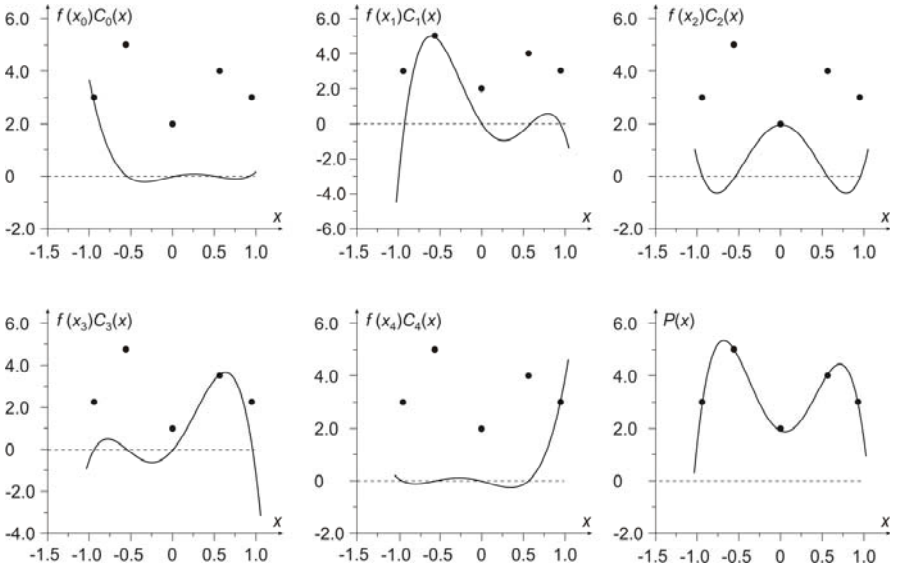

Fig. 4.6 Exemplary components of polynomial (4.31)

From (4.20), it can be noticed that shifted Legendre polynomials in the interval $[0, b]$ are presented by (4.32)

$$P_{k+1}(x) = \frac{1}{k+1}\left[(2k+1)\left(-1+\frac{2x}{b}\right) P_k\left(-1+\frac{2x}{b}\right) - k\, P_{k-1}\left(-1+\frac{2x}{b}\right) \right]$$
$$P_0(x) = 1 \tag{4.32}$$
$$P_1(x) = -1 + \frac{2x}{b}$$

A few shifted Legendre polynomials in the interval $[0, b]$ are given by

$$
\begin{aligned}
&P_0(x) = 1\\
&P_1(x) = -1 + 2x/b\\
&P_2(x) = 1 - 8x/b + 18x^2/b^2 - 12x^3/b^3\\
&P_3(x) = -16/9 + 16x/b - 188\,x^2/3b^2 + 1160\,x^3/9b^3 - 400x^4/3b^4 + 160x^5/3b^5\\
&P_4(x) = -139/36 + 863/18\,b - 4673x^2/18b^2 + 7310x^3/9b^3 - 14330x^4/9b^4\\
&\qquad + 17360x^5/9b^5 - 3920x^6/3b^6 + 1120x^7/3b^7\\
&P_5(x) = -1507/180 + 11677x/90b - 27481x^2/30b^2 + 34918x^3/9b^3\\
&\qquad - 490324x^4/45\,b^4 + 21040^5/b^5 - 83368x^6/3b^6 + 23968x^7/b^7\\
&\qquad - 12096x^8/b^8 + 2688x^9/b^9
\end{aligned}
\tag{4.33}
$$

4.4 Hermite Polynomials

In each measuring point, Hermite polynomials $H(x)$ satisfy the conditions related to the individual measuring points and to the value of derivatives in these points

$$H(x_i) = f(x_i), \quad \frac{dH}{dx}(x_i) = \frac{d}{dx} f(x_i) \tag{4.34}$$

The individual Hermite polynomials can be determined with the use of the recurrence formula

$$
\begin{aligned}
&H_{k+1}(x) = 2xH_k(x) - 2kH_{k-1}(x)\\
&H_0(x) = 1\\
&H_1(x) = 2x, \quad k = 1, 2, \ldots, n-1
\end{aligned}
\tag{4.35}
$$

Some of the initial Hermite polynomials are as follows

$$
\begin{aligned}
H_0(x) &= 1\\
H_1(x) &= 2x\\
H_2(x) &= 4x^2 - 2\\
H_3(x) &= 8x^3 - 12x\\
H_4(x) &= 16x^4 - 48x^2 + 12\\
H_5(x) &= 32x^5 - 160x^3 + 120x
\end{aligned}
\tag{4.36}
$$

and are shown in Fig. 4.7.

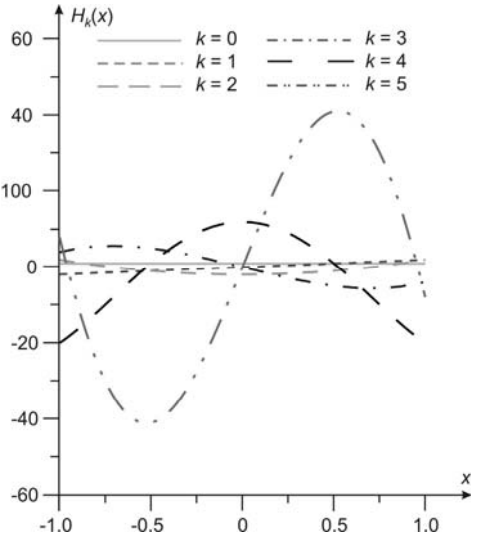

Fig. 4.7 The first six Hermite polynomials

Hermite polynomial $H(x)$ can be defined with a use of the cardinal functions $C_k(x)$ (4.8)

$$H(x) = \sum_{k=0}^{n-1}[H_k(x)f(x_k) + K_k(x)\frac{d}{dx}f(x)] \tag{4.37}$$

where

$$\begin{aligned} H_k(x) &= {C_k}^2(x)\left[1 - 2\frac{d}{dx}C_k(x_k)(x - x_k)\right] \\ K_k(x) &= {C_k}^2(x)(x - x_k) \end{aligned} \tag{4.38}$$

It is easy to see that the functions $H_k(x_i)$ and $K_k(x_i)$ fulfill the following relation

$$H_k(x_i) = \begin{cases} 1 & \text{if} \quad i = k \\ 0 & \text{if} \quad i \neq k \end{cases} \qquad \frac{d}{dx}H_k(x_i) = 0 \tag{4.39}$$

$$\frac{d}{dx}K_k(x_i) = \begin{cases} 1 & \text{if} \quad i = k \\ 0 & \text{if} \quad i \neq k \end{cases} \qquad K_k(x_i) = 0 \tag{4.40}$$

Hermite polynomials are orthogonal with the weight function $w(x) = e^{-x^2}$

$$\int_{-\infty}^{\infty} H_m(x)\, H_n(x)\, e^{-x^2}\, dx = \begin{cases} 0 & \text{if} \quad m \neq n \\ 2^n n!\, \sqrt{\pi} & \text{if} \quad m = n \end{cases} \tag{4.41}$$

Fig. 4.8 shows the components of polynomial (4.37) for five exemplary measuring points

$$[x, f(x)] = [1,\ 3;\ 2,\ 5;\ 3,\ 2;\ 4,\ 4;\ 5,\ 3] \text{ and } \frac{d}{dx} f(x) = [12.3;\ -3.7;\ -0.7;\ 3.3;\ -9.7]$$

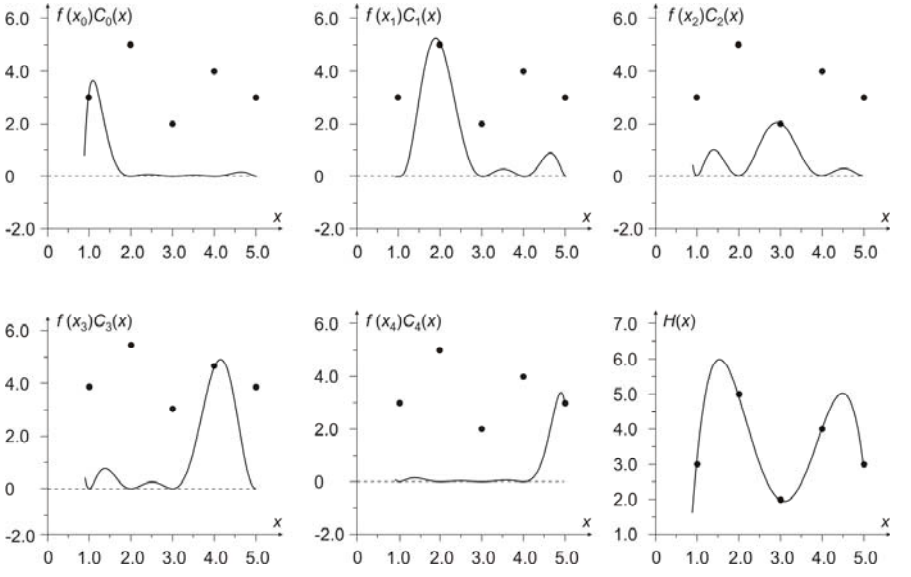

Fig. 4.8 Exemplary components of polynomial (4.36)

4.5 Cubic Splines

Cubic splines method is based on splitting the given interval into a collection of subintervals and constructing different approximating polynomials $S_k(x)$ at each subinterval. We use such a cubic polynomial between each successive pair of points, and the polynomial has the continuous first and second-order derivatives at these points.

We have three types of cubic splines, and the selection depends on the value of the second-order derivatives $S_0''(x_0)$ and $S_n''(x_n)$ at the end-points x_0 and x_n.

1. The natural spline, for which the second-order derivatives at the end-points equal zero, i.e. $S_0''(x_0) = S_n''(x_n) = 0$
2. The parabolic runout spline, for which the second-order derivatives at the first and second point are equal, i.e. $S_0''(x_0) = S_1''(x_1) \neq 0$. Regarding the last and one before last point, the second-order derivatives are equal and different then zero, i.e. $S''(x_n) = S''(x_{n-1}) \neq 0$
3. The cubic runout spline, for which the second-order derivatives at the end-points are different then zero and fulfill the following conditions $S_0''(x_0) = 2S_1''(x_1) - S_2''(x_2)$ and $S''(x_n) = 2S_{n-1}''(x_{n-1}) - S_{n-2}''(x_{n-2})$.

The general form of cubic polynomial is as follows

$$S_k(x) = a_k + b_k(x - x_k) + c_k(x - x_k)^2 + d_k(x - x_k)^3 \tag{4.42}$$

For each measuring points x_k we have

$$S_k(x_k) = f(x_k), \quad k = 0, 1, ..., n-1 \tag{4.43}$$

where *n* is the number of measuring points

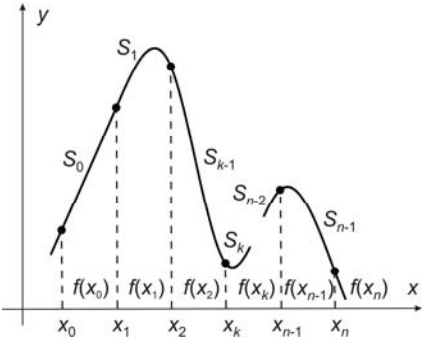

Fig. 4.9 Cubic splines

For each point, except the first and the last point, the particular polynomials fulfill the following conditions

$$S_k(x_k) = S_{k-1}(x_k) \tag{4.44}$$

$$S_k'(x_k) = S_{k-1}'(x_k) \tag{4.45}$$

$$S_k''(x_k) = S_{k-1}''(x_k) \tag{4.46}$$

In order to determine unknown coefficients a_k, b_k, c_k, d_k let us use Eqs.(4.42 – 4.46). Thus we have

$$S_{k-1}(x_k) = a_{k-1} + b_{k-1}(x_k - x_{k-1}) + c_{k-1}(x_k - x_{k-1})^2 + d_{k-1}(x_k - x_{k-1})^3 \tag{4.47}$$

and

$$S_k(x_k) = S_{k-1}(x_k) = a_k = f(x_k) \tag{4.48}$$

Denoting the difference between successive points by Δ

$$\Delta = x_k - x_{k-1} \tag{4.49}$$

Eq. (4.47) becomes

$$S_{k-1}(x_k) = a_{k-1} + b_{k-1}\Delta + c_{k-1}\Delta^2 + d_{k-1}\Delta^3 = a_k = f(x_k) \tag{4.50}$$

Taking Eq. (4.45) into consideration, we obtain

$$S'_{k-1}(x_k) = b_{k-1} + 2c_{k-1}(x_k - x_{k-1}) + 3d_{k-1}(x_k - x_{k-1})^2 \tag{4.51}$$

and because

$$S'_k(x_k) = b_k \tag{4.52}$$

hence

$$b_k + 2c_k\Delta + 3d_k\Delta^2 = b_{k+1} \tag{4.53}$$

In a similar way, on the basis of Eq. (4.46) we have

$$S''_{k-1}(x_i) = 2c_{k-1} + 6d_{k-1}\Delta \tag{4.54}$$

$$S''_k(x_k) = 2c_k \tag{4.55}$$

and

$$2c_{k-1} + 6d_{k-1}\Delta = 2c_k \tag{4.56}$$

hence

$$d_{k-1} = \frac{2c_k - 2c_{k-1}}{6\Delta} \tag{4.57}$$

and

$$d_k = \frac{2c_{k+1} - 2c_k}{6\Delta} = \frac{M_{k+1} - M_k}{6\Delta} \tag{4.58}$$

where

$$2c_k = M_k \tag{4.59}$$

From Eq. (4.50), we obtain

$$b_k = \frac{a_{k+1} - a_k}{\Delta} - \Delta(d_k \Delta + c_k) \tag{4.60}$$

Taking the relations (4.48) and (4.58) into account in the formula (4.60), we get

$$b_k = \frac{y_{k+1} - y_k}{\Delta} - \Delta \frac{M_{k+1} + 2M_k}{6} \tag{4.61}$$

Substituting the relations (4.59) and (4.61) into Eq. (4.53), we obtain

$$\begin{gathered} M_k + 4M_{k+1} + M_{k+2} = \frac{6}{\Delta^2}\left[f(x_k) - 2f(x_{k+1}) + f(x_{k+2})\right] \\ k = 0,1,...,n-2 \end{gathered} \tag{4.62}$$

Eq. (4.62) can be represented by the following matrix equation

$$\begin{bmatrix} 1 & 4 & 1 & 0 & . & . & 0 & 0 & 0 & 0 \\ 0 & 1 & 4 & 1 & . & . & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 4 & . & . & 0 & 0 & 0 & 0 \\ . & . & . & . & . & . & . & . & . & . \\ 0 & 0 & 0 & 0 & . & . & 4 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & . & . & 1 & 4 & 1 & 0 \\ 0 & 0 & 0 & 0 & . & . & 0 & 1 & 4 & 1 \end{bmatrix} \cdot \begin{bmatrix} M_0 \\ M_1 \\ M_2 \\ . \\ M_{n-2} \\ M_{n-1} \\ M_n \end{bmatrix} = \frac{6}{\Delta^2} \begin{bmatrix} f(x_0) - 2f(x_1) + f(x_2) \\ f(x_1) - 2f(x_2) + f(x_3) \\ f(x_2) - 2f(x_3) + f(x_4) \\ . \\ . \\ f(x_{n-3}) - 2f(x_{n-2}) + f(x_{n-1}) \\ f(x_{n-2}) - 2f(x_{n-1}) + f(x_n) \end{bmatrix} \tag{4.63}$$

The solution of the above equations with respect to $M_0 - M_n$ enables to determine the unknown coefficients of the polynomial (4.42) on the basis of Eqs. (4.48), (4.58), (4.59) and (4.61).

Eq. (4.63) can be reduced to the one of the three forms, in relation to the type of splines i.e. the natural splines, the parabolic runout splines and the cubic runout splines.

For the natural splines, on the basis of Eqs. (4.55) and (4.59) we have

$$M_1 = M_n = 0 \tag{4.64}$$

hence, in the left hand side of Eq. (4.63), the first and last column of the matrix can be eliminated and the equation can be rewritten to the form

$$\begin{bmatrix} 4 & 1 & 0 & 0 & . & . & 0 & 0 & 0 & 0 \\ 1 & 4 & 1 & 0 & . & . & 0 & 0 & 0 & 0 \\ 0 & 1 & 4 & 1 & . & . & 0 & 0 & 0 & 0 \\ . & . & . & . & . & . & . & . & . & . \\ 0 & 0 & 0 & 0 & . & . & 1 & 4 & 1 & 0 \\ 0 & 0 & 0 & 0 & . & . & 0 & 1 & 4 & 1 \\ 0 & 0 & 0 & 0 & . & . & 0 & 0 & 1 & 4 \end{bmatrix} \cdot \begin{bmatrix} M_1 \\ M_2 \\ M_3 \\ . \\ M_{n-3} \\ M_{n-2} \\ M_{n-1} \end{bmatrix} = \frac{6}{\Delta^2} \begin{bmatrix} f(x_0) - 2f(x_1) + f(x_2) \\ f(x_1) - 2f(x_2) + f(x_3) \\ f(x_2) - 2f(x_3) + f(x_4) \\ . \\ . \\ f(x_{n-3}) - 2f(x_{n-2}) + f(x_{n-1}) \\ f(x_{n-2}) - 2f(x_{n-1}) + f(x_n) \end{bmatrix} \tag{4.65}$$

Fig. 4.10 shows the natural spline. Implementing the condition (4.64) in effect makes the cubic function outside the end-points pass into a straight line.

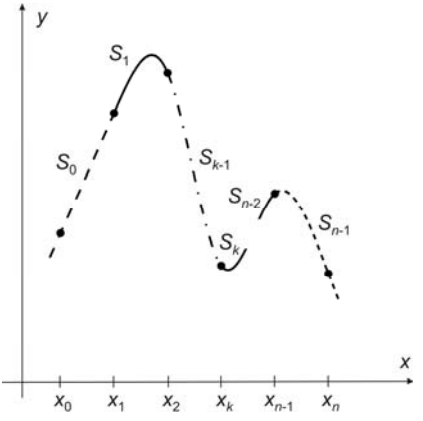

Fig. 4.10 Natural spline

For the parabolic runout spline, for which

$$M_0 = M_1 \tag{4.66}$$

$$M_n = M_{n-1} \tag{4.67}$$

Eq. (4.63) can be simplified to the following form

$$\begin{bmatrix} 5 & 1 & 0 & 0 & . & . & 0 & 0 & 0 & 0 \\ 1 & 4 & 1 & 0 & . & . & 0 & 0 & 0 & 0 \\ 0 & 1 & 4 & 1 & . & . & 0 & 0 & 0 & 0 \\ . & . & . & . & . & . & . & . & . & . \\ 0 & 0 & 0 & 0 & . & . & 1 & 4 & 1 & 0 \\ 0 & 0 & 0 & 0 & . & . & 0 & 1 & 4 & 1 \\ 0 & 0 & 0 & 0 & . & . & 0 & 0 & 1 & 5 \end{bmatrix} \cdot \begin{bmatrix} M_1 \\ M_2 \\ M_3 \\ . \\ M_{n-3} \\ M_{n-2} \\ M_{n-1} \end{bmatrix} = \frac{6}{\Delta^2} \begin{bmatrix} f(x_0) - 2f(x_1) + f(x_2) \\ f(x_1) - 2f(x_2) + f(x_3) \\ f(x_2) - 2f(x_3) + f(x_4) \\ . \\ . \\ f(x_{n-3}) - 2f(x_{n-2}) + f(x_{n-1}) \\ f(x_{n-2}) - 2f(x_{n-1}) + f(x_n) \end{bmatrix} \tag{4.68}$$

Fig. 4.11 presents the parabolic runout spline. Implementing the condition (4.67) and (4.68) in effect makes the cubic function outside the end-points pass into a parabola.

For the cubic runout spline, we have

$$M_0 = 2M_1 - M_2 \tag{4.69}$$

$$M_n = 2M_{n-1} - M_{n-2} \tag{4.70}$$

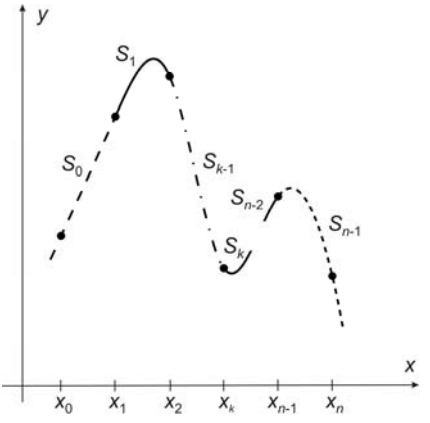

Fig. 4.11 Parabolic runout spline

and now Eq. (4.63) takes the form

$$\begin{bmatrix} 6 & 0 & 0 & 0 & . & . & 0 & 0 & 0 & 0 \\ 1 & 4 & 1 & 0 & . & . & 0 & 0 & 0 & 0 \\ 0 & 1 & 4 & 1 & . & . & 0 & 0 & 0 & 0 \\ . & . & . & . & . & . & . & . & . & . \\ 0 & 0 & 0 & 0 & . & . & 1 & 4 & 1 & 0 \\ 0 & 0 & 0 & 0 & . & . & 0 & 1 & 4 & 1 \\ 0 & 0 & 0 & 0 & . & . & 0 & 0 & 0 & 6 \end{bmatrix} \cdot \begin{bmatrix} M_1 \\ M_2 \\ M_3 \\ . \\ M_{n-3} \\ M_{n-2} \\ M_{n-1} \end{bmatrix} = \frac{6}{\Delta^2} \begin{bmatrix} f(x_0) - 2f(x_1) + f(x_2) \\ f(x_1) - 2f(x_2) + f(x_3) \\ f(x_2) - 2f(x_3) + f(x_4) \\ . \\ . \\ f(x_{n-3}) - 2f(x_{n-2}) + f(x_{n-1}) \\ f(x_{n-2}) - 2f(x_{n-1}) + f(x_n) \end{bmatrix} \tag{4.71}$$

Fig. 4.12 presents the cubic runout spline, for which the cubic function outside of the end-points does not pass into any other function

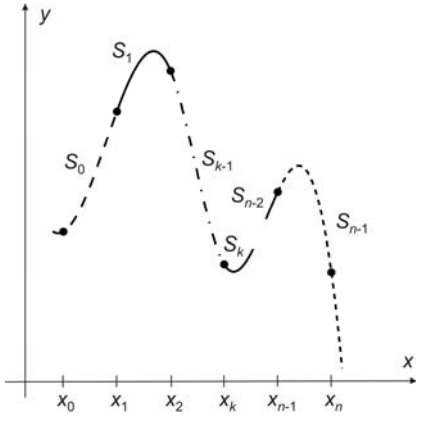

Fig. 4.12 Cubic runout spline

4.6 The Least-Squares Approximation

The approximation of the measuring data by means of Lagrange, Tchebyshev, Legendre and Hermite polynomials leads to the derivation of a polynomial of the order equal to the number of approximation points. For a large number of such points, the derived polynomial would be then of a very high order. In such a situation, it is better in many cases to construct a relatively low order polynomial, which is passing close to the measuring data instead of cutting across them. In the method of the least-squares approximation, the polynomial is such that the sum of squares of the differences between the ordinates of the approximation line and the measuring points is at minimum

$$\sum_{k=0}^{n}\left[f(x_k)-Q(x_k)\right]^2=\min. \tag{4.72}$$

Let the polynomial of the degree $m<n$ have the form

$$Q(x)=\sum_{i=0}^{m}a_i x^i \tag{4.73}$$

For a minimum (4.72) with respect to the parameters a_i for $i=0,1,...,m,$ it is necessary that

$$\frac{\partial Q(x)}{\partial a_0}=\frac{\partial Q(x)}{\partial a_1}=...=\frac{\partial Q(x)}{\partial a_m}=0 \tag{4.74}$$

Let us present Eq. (4.72) as follows

$$\sum_{k=0}^{n}f^2(x_k)-2\sum_{k=0}^{n}Q(x_k)f(x_k)+\sum_{k=0}^{n}Q(x_k)^2=\min. \tag{4.75}$$

Substituting the expression for $Q(x_k)$ (4.73) into the left-hand side of (4.75) gives

$$\sum_{k=0}^{n}f^2(x_k)-2\sum_{k=0}^{n}\left(\sum_{i=0}^{m}a_i x_k^i\right)f(x_k)+\sum_{k=0}^{n}\left(\sum_{i=0}^{m}a_i x_k^i\right)^2=\min. \tag{4.76}$$

A simple calculation of the derivatives, according to (4.74), leads (4.76) to the following system of equations denoted in the normal form

$$\mathbf{X}\,\mathbf{A} = \mathbf{Y}$$

where

$$\mathbf{X} = \begin{bmatrix} \sum_{k=0}^{n} x_k^0 & \sum_{k=0}^{n} x_k^1 & \sum_{k=0}^{n} x_k^2 & . & \sum_{k=0}^{n} x_k^m \\ \sum_{k=0}^{n} x_k^1 & \sum_{k=0}^{n} x_k^2 & \sum_{k=0}^{n} x^3 & . & \sum_{k=0}^{n} x_k^{m+1} \\ . & . & . & . & . \\ . & . & . & . & . \\ \sum_{k=0}^{n} x_k^m & \sum_{k=0}^{n} x_k^{m+1} & \sum_{k=0}^{n} x_k^{m+2} & . & \sum_{k=0}^{n} x_k^{2m} \end{bmatrix} \tag{4.77}$$

$$\mathbf{A} = \begin{bmatrix} a_0 \\ a_1 \\ . \\ . \\ a_m \end{bmatrix} \qquad \mathbf{Y} = \begin{bmatrix} \sum_{k=0}^{n} f(x_k) x_k^0 \\ \sum_{k=0}^{n} f(x_k) x_k^1 \\ . \\ . \\ \sum_{k=0}^{n} f(x_k) x_k^m \end{bmatrix}$$

a solution of which is given by

$$\mathbf{A} = [\mathbf{X}^T \mathbf{X}]^{-1} \mathbf{X}^T \mathbf{Y} \tag{4.78}$$

Note that the success of the approximation developed by the least squares method depends very much on the accuracy of all intermediate calculations. For this reason, the calculations should be carried out with maximum possible precision and the necessary rounding up should be limited to a minimum.

4.7 Relations between Coefficients of the Models

Let coefficients a_k of the polynomial $M(x)$

$$M(x) = \sum_{k=0}^{n} a_k x^k \tag{4.79}$$

be equal to

$$a_k = \frac{1}{k!} A_k \tag{4.80}$$

Let additionally A_k represents Maclaurin series coefficients, hence they are the successive derivatives of $M(x)$ for $x = 0$

$$A_0 = M(x)\big|_{x=0}, \quad A_1 = \frac{dM(x)}{dx}\big|_{x=0}, \quad A_2 = \frac{d^2M(x)}{dx^2}\big|_{x=0},$$
$$..., \; A_k = \frac{d^kM(x)}{dx^k}\big|_{x=0} \tag{4.81}$$

The mutual relations between the coefficients A_k (4.81) and the coefficients $a_0, a_1, ..., a_{n-1}$ and $b_0, b_1, ..., b_m$ of a Laplace transfer function

$$K(s) = \frac{Y(s)}{U(s)} = \frac{b_m s^m + b_{m-1}s^{m-1} + ... + b_1 s + b_0}{s^n + a_{n-1}s^{n-1} + ... + a_1 s + a_0} \tag{4.82}$$

or the state equation

$$\begin{aligned} \dot{\mathbf{x}}(t) &= \mathbf{A}\mathbf{x}(t) + \mathbf{B}u(t), \quad \mathbf{x}(0) = 0 \\ y(t) &= \mathbf{C}\mathbf{x}(t) \end{aligned} \tag{4.83}$$

where $\mathbf{x}(t)$ is the state vector, **A**, **B** and **C** are the real matrices of corresponding dimensions

$$\mathbf{A} = \begin{bmatrix} 0 & 1 & 0 & . & 0 \\ 0 & 0 & 1 & . & 0 \\ . & . & . & . & . \\ 0 & 0 & 0 & . & 1 \\ -a_0 & -a_1 & . & . & -a_{n-1} \end{bmatrix} \quad \mathbf{B} = \begin{bmatrix} 0 \\ 0 \\ . \\ . \\ 1 \end{bmatrix} \quad \mathbf{C} = \begin{bmatrix} b_0 & b_1 & . & . & b_m \end{bmatrix} \tag{4.84}$$

are given by the following matrix equation

$$\begin{bmatrix} b_m \\ b_{m-1} \\ . \\ . \\ b_0 \\ a_{n-1} \\ a_{n-2} \\ . \\ . \\ a_0 \end{bmatrix} = \begin{bmatrix} \begin{pmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{pmatrix} & \begin{pmatrix} 0 & 0 & . & 0 & 0 \\ -A_0 & 0 & . & 0 & 0 \\ -A_1 & -A_0 & . & 0 & 0 \\ . & . & . & . & . \\ -A_{n-2} & -A_{n-3} & . & -A_0 & 0 \end{pmatrix} \\ \begin{pmatrix} 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{pmatrix} & \begin{pmatrix} -A_{n-1} & . & . & -A_1 & -A_0 \\ -A_n & . & . & -A_2 & -A_1 \\ . & . & . & . & . \\ . & . & . & . & . \\ -A_{2n-2} & . & . & -A_n & -A_{n-1} \end{pmatrix} \end{bmatrix}^{-1} \begin{bmatrix} A_0 \\ A_1 \\ A_2 \\ . \\ . \\ . \\ . \\ . \\ A_{2n-1} \end{bmatrix} \tag{4.85}$$

For the first three values of n, the equation (4.85) is reduced to the following form:

for $n = 1$

$$\begin{bmatrix} b_0 \\ a_0 \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 0 & -A_0 \end{bmatrix}^{-1} \begin{bmatrix} A_0 \\ A_1 \end{bmatrix} \tag{4.86}$$

for $n = 2$

$$\begin{bmatrix} b_1 \\ b_0 \\ a_1 \\ a_0 \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & -A_0 & 0 \\ 0 & 0 & -A_1 & -A_0 \\ 0 & 0 & -A_2 & -A_1 \end{bmatrix}^{-1} \begin{bmatrix} A_0 \\ A_1 \\ A_2 \\ A_3 \end{bmatrix} \tag{4.87}$$

and for $n = 3$

$$\begin{bmatrix} b_2 \\ b_1 \\ b_0 \\ a_2 \\ a_1 \\ a_0 \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & -A_0 & 0 & 0 \\ 0 & 0 & 1 & -A_1 & -A_0 & 0 \\ 0 & 0 & 0 & -A_2 & -A_1 & -A_0 \\ 0 & 0 & 0 & -A_3 & -A_2 & -A_1 \\ 0 & 0 & 0 & -A_4 & -A_3 & -A_2 \end{bmatrix}^{-1} \begin{bmatrix} A_0 \\ A_1 \\ A_2 \\ A_3 \\ A_4 \\ A_5 \end{bmatrix} \tag{4.88}$$

The reverse relation is given by Eq. 4.89.

$$\begin{bmatrix} A_0 \\ A_1 \\ A_2 \\ \cdot \\ \cdot \\ \cdot \\ \cdot \\ A_{2n-1} \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 & 0 & \cdot & 0 & 0 & 0 \\ a_{n-1} & 1 & 0 & 0 & \cdot & 0 & 0 & 0 \\ a_{n-2} & a_{n-1} & 1 & 0 & \cdot & 0 & 0 & 0 \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ a_0 & a_1 & a_2 & a_3 & \cdot & 0 & 0 & 0 \\ 0 & a_0 & a_1 & a_2 & \cdot & 1 & 0 & 0 \\ 0 & 0 & a_0 & a_1 & \cdot & a_{n-1} & 1 & 0 \\ 0 & 0 & 0 & a_0 & \cdot & a_{n-2} & a_{n-1} & 1 \end{bmatrix}^{-1} \begin{bmatrix} b_m \\ b_{m-1} \\ b_{m-2} \\ \cdot \\ b_0 \\ 0 \\ \cdot \\ 0 \end{bmatrix} \tag{4.89}$$

It permits to calculate the coefficients $A_0, A_1, ..., A_{2n-1}$ of the Maclaurin series having knowledge of parameters $a_0,\ a_1, ..., a_{n-1},\ b_0,\ b_1, ..., b_m$.

Note that the subsequent coefficients of the series $A_{2n}, A_{2n+1}, ..., A_{2n+2},$ of the first column in equation (4.89), are expressed by the coefficients $A_0, A_1, ..., A_{2n-1}$ preceding them. The relations between the discussed coefficients are shown below

for $n = 1$

$$\begin{bmatrix} A_2 \\ A_3 \\ A_4 \\ . \\ . \end{bmatrix} = \begin{bmatrix} -A_1 \\ -A_2 \\ -A_3 \\ . \\ . \end{bmatrix} a_0 \tag{4.90}$$

for $n = 2$

$$\begin{bmatrix} A_4 \\ A_5 \\ A_6 \\ . \\ . \end{bmatrix} = \begin{bmatrix} -A_3 & -A_2 \\ -A_4 & -A_3 \\ -A_5 & -A_4 \\ . & . \\ . & . \end{bmatrix} \begin{bmatrix} a_1 \\ a_0 \end{bmatrix} \tag{4.91}$$

and for $n = 3$

$$\begin{bmatrix} A_6 \\ A_7 \\ A_8 \\ . \\ . \end{bmatrix} = \begin{bmatrix} -A_5 & -A_4 & -A_3 \\ -A_6 & -A_5 & -A_4 \\ -A_7 & -A_6 & -A_5 \\ . & . & . \\ . & . & . \end{bmatrix} \begin{bmatrix} a_2 \\ a_1 \\ a_0 \end{bmatrix} \tag{4.92}$$

From $(4.80) - (4.92)$ it can be seen that $2n$ initial coefficients of the power series expansion contain all the information describing polynomial $M(x)$.

It is important to note that the application of the Maclaurin series allow models development which is particularly useful in the case of systems operating in dynamic states. That is because the series refers to the functions defined near the origin.

4.8 Standard Nets

When an object is under non-parametric identification procedure, in many cases it is essential to know the order of its model.

During the identification procedure, we use a wide range of different methods selecting these, which are most suitable for the type of an object under identification. The lack of any universal identification method on the one hand, and a large variety of objects on the other implicate serious problems with choosing the correct method of identification. Thus a great amount of afford is required to obtain a correct final effect.

As an example, let us consider the three most common groups of objects and the methods applied during the identification process

1. Inertial objects, which are identified through the analysis of the step-response ordinates
2. A class of oscillatory objects, for which a number of methods is applied, like the two consecutive extremes method, readings of the step-response ordinates, the method of apparent move of zeros and poles of a transfer function
3. Multi-inertial objects of the order denoted by the integer or fractions. These are identified either through the analysis of the initial interval of the step-response or by means of the finite difference method with the use of the auxiliary function to determine a rank and type of inertia. Using one of these two methods, it is possible to reduce the transfer function of multi-inertial objects to the Strejc model. The latter is particularly useful to present object dynamics with step characteristics increasing monotonically.

Summarizing, each group of objects is identified in a different way. A number of various methods can be used for this aim. In the following pages, we present the universal solution, to some degree, of the parametric identification problem. It is based on the standard nets method and computer math-programs like MATLAB, Maple, MathCad and LabVIEW.

The central point of the method is a comparison of identification nets. The standard identification nets are compared with the identification net of an object under modelling. If initial parts of the nets characteristics are compatible, it permits to determine the type of the object model. It corresponds with the model, for which the standard identification net has been selected.

The standard identification nets are determined most often for the following 13 models presented below by the formulae (4.93) – (4.105)

$$K(s) = \prod_{i=1}^{n} \frac{k_i}{(1 + sT_i)} \tag{4.93}$$

$$K(s) = \prod_{i=1}^{n} \frac{k_i}{(1 + sT_i)} (1 + sT_{n+1}) \tag{4.94}$$

$$K(s) = \prod_{i=1}^{n} \frac{k_i}{(1 + sT_i)} e^{-s\tau} \tag{4.95}$$

$$K(s) = \prod_{i=1}^{n} \frac{k_i}{(1 + sT_i)} (1 - e^{-s\tau}) \tag{4.96}$$

$$K(s) = \prod_{i=1}^{n} \frac{k_i}{(1 + sT_i)} sT \tag{4.97}$$

$$K(s) = \frac{1 + \dfrac{2\beta}{\omega_0} s}{\dfrac{s^2}{{\omega_0}^2} + \dfrac{2\beta}{\omega_0} s + 1} \tag{4.98}$$

$$K(s) = \frac{\dfrac{k}{{\omega_0}^2}}{\dfrac{s^2}{{\omega_0}^2} + 2\beta\dfrac{s}{\omega_0} + 1} \tag{4.99}$$

$$K(s) = \frac{1}{sT} \tag{4.100}$$

$$K(s) = \frac{1}{sT_1}\frac{1}{(1+T_2 s)} \tag{4.101}$$

$$K(s) = sT \tag{4.102}$$

$$K(s) = \frac{s}{s^2 + {\omega_0}^2} \tag{4.103}$$

$$K(s) = \frac{kTs}{(1+Ts)\left(\dfrac{s^2}{{\omega_0}^2} + \dfrac{2\beta}{\omega_0}s + 1\right)} \tag{4.104}$$

$$K(s) = \frac{k}{(1+s\overline{T})^n}, \qquad \overline{T} = \frac{T}{n} \tag{4.105}$$

Step-responses are used for the development of standard identification nets, and for all listed models, they can be easily obtained applying the inverse Laplace transform. It is only the model (4.105), which may produce some difficulties. Its step-response is

$$h(t) = \mathcal{L}^{-1}\left(\frac{k}{s(1+s\overline{T})^n}\right) \tag{4.106}$$

and in the time-domain is

$$h(t) = k\left(1 - \sum_{m=1}^{n} \frac{1}{\overline{T}^{m-1}(m-1)!} t^{m-1} \exp\left(\frac{-t}{\overline{T}}\right)\right) \tag{4.107}$$

For a fractional n, this way of calculations is not possible. However, in such a case, the response $h(t)$ can be determined using Gamma Euler functions $\Gamma(n)$ and $\Gamma(n, t/\overline{T})$. Hence, we have

$$h(t) = k\left(1 - \frac{\Gamma(n, t/\overline{T})}{\Gamma(n)}\right) \tag{4.108}$$

where

$$\Gamma(n) = \int_0^{\infty} t^{n-1} e^{-t} dt \tag{4.109}$$

and

$$\Gamma(n, t/\overline{T}) = \int_{t/\overline{T}}^{\infty} t^{n-1} e^{-t} dt \tag{4.110}$$

The standard identification net is obtained through the transform of $h(t)$ response using the parametric equations (4.111) and (4.112),

$$X(t) = f_1\left(\phi(t),\ \phi(t/a)\right) \tag{4.111}$$

$$Y(t) = f_1\left(\phi(t),\ \phi(t/a)\right) \tag{4.112}$$

The coordinates $X(t)$ (4.111) and $Y(t)$ (4.112) are calculated using any of the three algorithms presented by the formulae (4.113) – (4.115).

$$X(t) = \frac{\phi(t) + \phi(t/a)}{2} \qquad Y(t) = \frac{\phi_n(t) - \phi_n(t/a)}{2} \tag{4.113}$$

or

$$X(t) = \frac{\phi(t/a)}{\phi(t) + c} \qquad Y(t) = \sqrt{\left|\phi(t)\phi(t/a)\right|} \tag{4.114}$$

and

$$X(t) = \frac{\phi(t) - \phi(t/a)}{\left(\sqrt{\phi(t/a)} + \sqrt{\phi(t)}\right)^2 + c} \qquad Y(t) = \frac{\phi(t) + \phi(t/a)}{\left(\sqrt{\phi(t/a)} + \sqrt{\phi(t)}\right)^2 + c} \tag{4.115}$$

where

$$\phi(t) = \frac{h(t) - h(0)}{h(\infty) - h(0)} \tag{4.116}$$

$$\phi(t/a) = \frac{h(t/a) - h(0)}{h(\infty) - h(0)} \tag{4.117}$$

In Eqs. (4.113) – (4.127), a is the parameter related to a number of samples of the digitized step-response $h(t)$. The optimum solution can be obtained for $a = 2$. The infinitesimal $c \in \Re$ protects the denominator from being equal zero.

It is convenient to group the standard identification nets according the class of objects: multi-inertial, multi-inertial with a delay and oscillatory nets.

Fig. 4.13 and Fig. 4.14 show the initial parts of exemplary families of the standard identification nets, for two models (4.99) and (4.105). They are obtained

through the identification algorithm (4.113) for $k=1,\ \omega_0=1$ and $\beta=0.1,\ 0.3,\ 0.5,\ 0.7,\ 0.9,\ t\in[0,15]$ for the model (4.99), and for $a=2,\ k=1,\ \overline{T}=1,\ n=1,\ 2,...,5,\ t\in[0,\ 15]$ for the model (4.105).

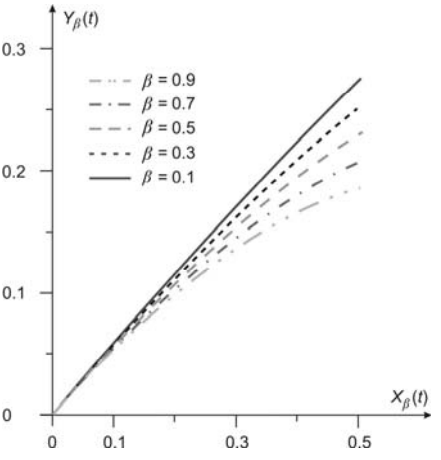

Fig. 4.13 Family of standard identification nets for model (4.99) $a=2,\ k=1,\ \omega_0=1,\ t\in[0,15]$

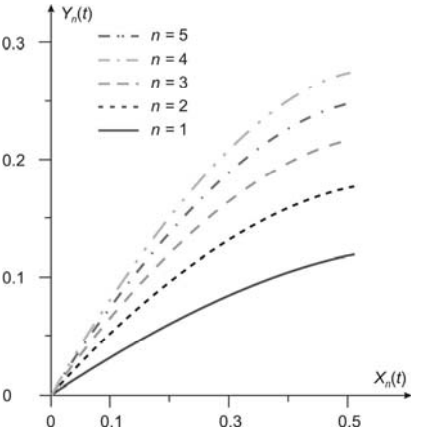

Fig. 4.14 Family of standard identification nets for model (4.105), $a=2,\ k=1,\ \overline{T}=1,\ t\in[0,15]$

The construction of nets Fig. 4.13 and Fig. 4.14 is based on measurements and data of step-responses. The latter can easily be transformed into identification nets using the formulae (4.113) – (4.115). The whole process can be carried out fully

automatically through the application of the measuring system shown in Fig.1.1, and additionally supported by special software tools for measurement and control. These requirements are satisfied in the best way by LabVIEW software.

For inertial object of class (4.105), it is practically convenient to apply the graph $\max(X_n(t), Y_n(t))$ shown in Fig. 4.15. This way allows for an easy estimation of the fractional order of inertia. For $n = 1, 2, ..., 5,$ such a graph is shown in Fig.4.16.

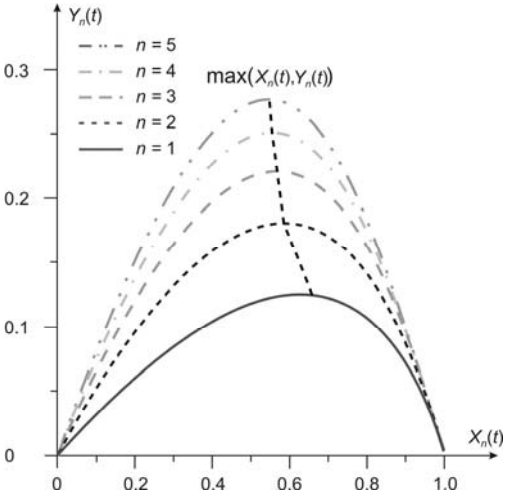

Fig. 4.15 Maximum points for model (4.105)

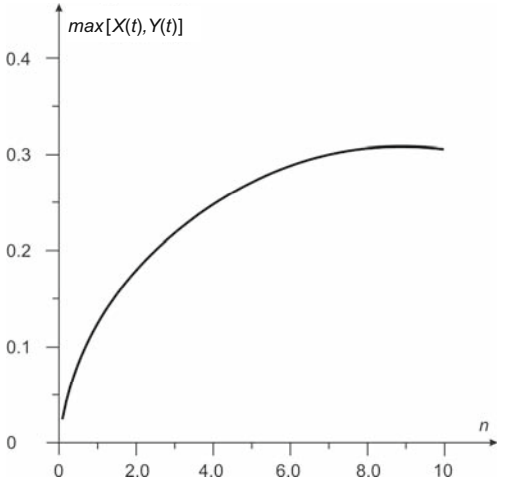

Fig. 4.16 Maximum of standard identification nets n – inertia order of model (4.105)

4.9 Levenberg-Marquardt Algorithm

In this subchapter, we will present Levenberg-Marquardt optimization algorithm and discuss the potential of using it for identification. Application of this algorithm has many advantages in comparison with other optimization methods. It combines the steepest descent method with Gauss-Newton method, and operates correctly in search for parameters both far from and close to the optimum one. In the first situation the algorithm of the linear model of steepest descent is used, and in the second one-the squared convergence. The fast convergence is the additional advantage of the algorithm.

Levenberg-Marquardt algorithm is the iterative method, in which the vector of unknown parameters, for the step $k+1$, is determined by the equation

$$\mathbf{z}_{k+1} = \mathbf{z}_k{}^T [\mathbf{J}^T(\mathbf{z}_k, x)\mathbf{J}(\mathbf{z}_k, x) + \mu_k \mathbf{I}]^{-1} \mathbf{J}^T(\mathbf{z}_k, x)\varepsilon(\mathbf{z}_k, x) \tag{4.118}$$

with the approximation error

$$\varepsilon(\mathbf{z}_k, x) = \sum_{i=1}^{n} \varepsilon(\mathbf{z}_k, x_i)^2 \tag{4.119}$$

where

$$\varepsilon(\mathbf{z}_k, x) = \begin{bmatrix} y_1 - \phi(z_1, x) \\ y_2 - \phi(z_2, x) \\ . \\ y_n - \phi(z_n, x) \end{bmatrix} \tag{4.120}$$

$$\mathbf{J}(\mathbf{z}_k, x) = \begin{bmatrix} \frac{\partial \varepsilon(\mathbf{z}_k, x_1)}{\partial z_1} & \frac{\partial \varepsilon(\mathbf{z}_k, x_1)}{\partial z_2} & . & \frac{\partial \varepsilon(\mathbf{z}_k, x_1)}{\partial z_m} \\ \frac{\partial \varepsilon(\mathbf{z}_k, x_2)}{\partial z_1} & \frac{\partial \varepsilon(\mathbf{z}_k, x_2)}{\partial z_2} & . & \frac{\partial \varepsilon(\mathbf{z}_k, x_2)}{\partial z_m} \\ . & . & . & . \\ \frac{\partial \varepsilon(\mathbf{z}_k, x_n)}{\partial z_1} & \frac{\partial \varepsilon(\mathbf{z}_k, x_n)}{\partial z_2} & . & \frac{\partial \varepsilon(\mathbf{z}_k, x_n)}{\partial z_m} \end{bmatrix} \tag{4.121}$$

The notations in $(4.118)-(4.121)$ are as follows: $k = 1, 2, ..., p$, p – a number of iteration loops, $\mathbf{J}_{nxm}(\mathbf{z}_k, x)$ – Jacobian matrix, $\mathbf{I}_{mxm}$ – unit matrix, μ_k – scalar, $\mathbf{x} = [x_1, x_2, ..., x_n]$ – vector of input parameters, $\mathbf{y} = [y_1, y_2, ..., y_n]$ – vector of output parameters, $\hat{y} = \varphi(z, x)$ – predicted model, $\mathbf{z} = [z_1, z_2, ..., z_m]$ – unknown parameters.

Levenberg-Marquardt algorithm is used for computation in two steps:

Step 1
the initial values of the vector $\mathbf{z}_k$

- assume the initial value of the coefficient μ_k (e.g. $\mu_k = 0.1$)
- solve the matrix equations (4.120) and (4.121)
- calculate the value of error (4.119)
- determine the parameters of the vector $\mathbf{z}_{k+1}$, according to (4.118).

Step 2 and further steps

- update the values of the parameter vector for the model $\hat{y}$
- solve the matrix equations (4.120), (4.121) and (4.118)
- calculate the value of error (4.119)
- compare the values of error (4.119) for the step k and the step $k-1$.

If the result is $\varepsilon(\mathbf{z}_k, x) \geq \varepsilon(\mathbf{z}_{k-1}, x)$, multiply μ_k by the specified value $\lambda \in \Re$ (e.g. $\lambda = 10$) and return to the step 2. If the result is $\varepsilon(\mathbf{z}_k, x) < \varepsilon(\mathbf{z}_{k-1}, x)$ divide μ_k by the value λ and return to the step 1.

If in the consecutive steps a decreasing in the value of error (4.119) is very small and insignificant, we then finish the iteration process. We fix $\mu_k = 0$ and determine the final result for the parameter vector.

If the value of coefficient μ_k is high, it means that the solution is not satisfactory. The values of the parameter vector $\mathbf{z}$ are not optimum ones, and the value of error (4.119) is not at minimum level. At this point it can be assumed

$$\mathbf{J}^T(\mathbf{z}_k, x)\mathbf{J}(\mathbf{z}_k, x) << \mu_k \mathbf{I} \tag{4.122}$$

and this leads to the steepest descent method, for which we have

$$\mathbf{z}_{k+1} = \mathbf{z}_k{}^T - \frac{1}{\mu_k}\mathbf{J}^T(\mathbf{z}_k, x)\varepsilon(\mathbf{z}_k, x) \tag{4.123}$$

If the value of the coefficient μ_k is small, it means that the values of the vector $\mathbf{z}$ parameters are close to the optimum solution,
then

$$\mathbf{J}^T(\mathbf{z}_k, x)\mathbf{J}(\mathbf{z}_k, x) >> \mu_k \mathbf{I} \tag{4.124}$$

and Levenberg-Marquardt algorithm is reduced to Gauss-Newton method

$$\mathbf{z}_{k+1} = \mathbf{z}_k{}^T - [\mathbf{J}^T(\mathbf{z}_k, x)\mathbf{J}(\mathbf{z}_k, x)]^{-1}\mathbf{J}^T(\mathbf{z}_k, x)\varepsilon(\mathbf{z}_k, x) \tag{4.125}$$

The selection of the coefficient values μ_k and λ depends on assumed programs and selected software.

4.9.1 Implementing Levenberg-Marquardt Algorithm Using LabVIEW

It is convenient to deploy Levenberg-Marquardt algorithm with LabVIEW software. Fig. 4.17. presents the block diagram of the measuring system for determining any characteristics of investigated object in this program Measurement data given by vectors **x** and **y** are sent to the measuring system through its analogue output and are recorded into the text files (Write to Measurement File1 and Write to Measurement File2, respectively). These data are next in the *Curve Fitting* block approximated by means of Levenberg-Marquardt algorithm. Fig. 4.18. presents the diagram of the general data approximation system, while Fig. 4.19. illustrates the *Curve Fitting* approximation block adapted for the exemplary approximation of frequency characteristic of the third-order system (in *Non-linear model* window).

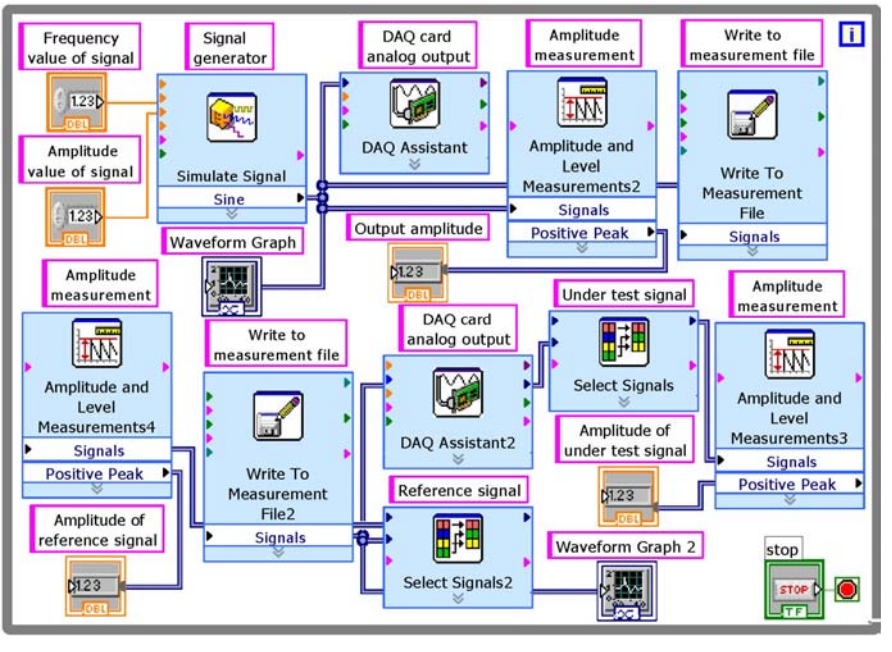

Fig. 4.17 Diagram of measuring system for determination of frequency characteristics

The approximation process is carried out for the initial value of parameters (*Initial guesses*) and number of iteration (*Maximum iterations*). Windows *Results* presents the value of calculated parameters and the value of mean square error (*MSE*).

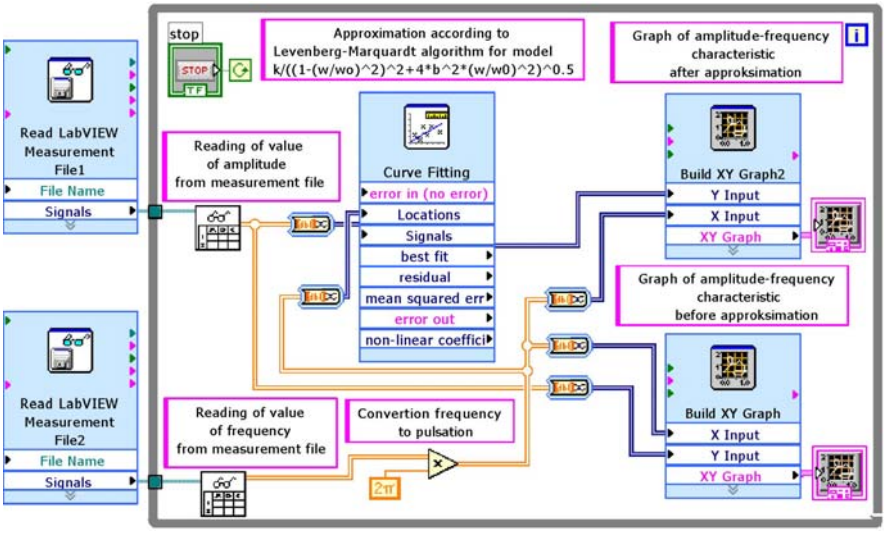

Fig. 4.18 Diagram of measuring system for approximation of frequency characteristics in LabVIEW

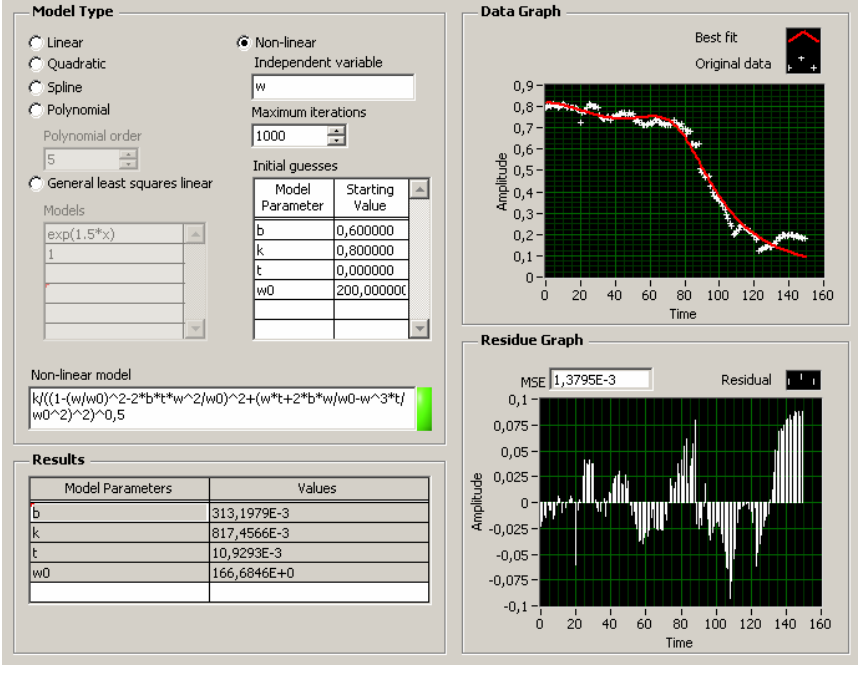

Fig. 4.19 Curve Fitting block (Fig. 4.18) for approximation of third-order system

4.10 Black-Box Identification

In the black-box identification, the experiment is carried out using discrete measurement data. From among preset parametric models, being a good match for these data, the desired model structure is selected. The discrete model of the identified object, in the form of the transfer function, is taken under consideration

$$K(z) = \frac{b_0 + b_1 z^{-1} + ... + b_m z^{-m}}{a_0 + a_1 z^{-1} + ... + z^{-n}} \tag{4.126}$$

or equivalent one

$$\mathbf{A}(z^{-1})\mathbf{y}[k] = \mathbf{B}(z^{-1})\mathbf{u}[k] \tag{4.127}$$

where

$$\mathbf{A}(z^{-1}) = a_0 + a_1 z^{-1} + ... + z^{-n} \tag{4.128}$$

$$\mathbf{B}(z^{-1}) = b_0 + b_1 z^{-1} + ... + b_m z^{-m} \tag{4.129}$$

$$z^{-n}\mathbf{x}[k] = \mathbf{x}[k-n], \quad n \in N \tag{4.130}$$

for which a white noise $\mathbf{e}[k]$ is added and the parametric model of ARX type (Auto Regressive with eXogenous input) is formulated.

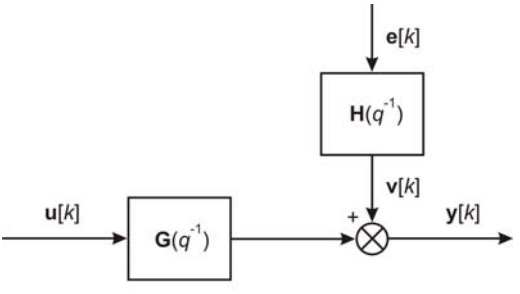

Fig. 4.20 ARX model, $\mathbf{v}(t)$ – noise

Now we have

$$\mathbf{A}(z^{-1})\mathbf{y}[k] = \mathbf{B}(z^{-1})\mathbf{u}[k] + \mathbf{e}[k] \tag{4.131}$$

and from it

$$\mathbf{y}[k] = \frac{\mathbf{B}(z^{-1})}{\mathbf{A}(z^{-1})}\mathbf{u}[k] + \frac{1}{\mathbf{A}(z^{-1})}\mathbf{e}[k] \tag{4.132}$$

where in (4.127) – (4.132) $\mathbf{u}[k]$, $\mathbf{y}[k]$ are the input and output signals at the discrete time k, $\mathbf{x}[k]$ is any measurement data in k and z^{-1} is delayed at one.

Denoting

$$\mathbf{H}(z^{-1}) = \frac{1}{\mathbf{A}(z^{-1})} \tag{4.133}$$

we finally have

$$\mathbf{y}[k] = \mathbf{K}(z^{-1})\mathbf{u}[k] + \mathbf{H}(z^{-1})\mathbf{e}[k] \tag{4.134}$$

Eq. (4.134) describes the ARX model shown in Fig. 4.20. Identification of the ARX model is based on the following assumptions

- the object $\mathbf{K}(z^{-1})$ is asymptotically stable
- the filter $\mathbf{H}(z^{-1})$ is linear, asymptotically stable, minimum-phase and invertible
- the input signal variation $\mathbf{u}[k]$ is sufficiently large

and leads to a simultaneous solution of two following tasks

- identification of the object, of which the transfer function is $\mathbf{K}(z^{-1})$
- identification of the filter $\mathbf{H}(q^{-1})$.

Let us present the equation (4.132) as follows

$$\mathbf{y}[k] = \mathbf{z}[k]\ \mathbf{\Phi} + \mathbf{e}[k] \tag{4.135}$$

where

$$\mathbf{z}[k] = [-\mathbf{y}[k-1], ..., -\mathbf{y}[k-1], \mathbf{u}[k], ..., \mathbf{u}[k-m]] \tag{4.136}$$

$$\mathbf{\Phi}^T = [a_0, ..., a_{n-1}, 1, b_0, ..., b_m] \tag{4.137}$$

Let us also denote by p the number of activating signals. Then Eq. (4.135) takes the final form

$$\begin{bmatrix} y_0[k] \\ . \\ y_{p-1}[k] \end{bmatrix} = \begin{bmatrix} z_0[k] \\ . \\ z_{p-1}[k] \end{bmatrix} \mathbf{\Phi} + \begin{bmatrix} e_0[k] \\ . \\ e_{p-1}[k] \end{bmatrix} \tag{4.138}$$

As the result the identification task is reduced to the determination of the estimates of model parameters

$$\hat{\mathbf{\Phi}} = \Theta\left(n,\ m,\ \mathbf{u}[k], \mathbf{y}[k]\right) \tag{4.139}$$

Apart of the ARX model, there are also other structures applied quite often:

– AR model described by the equation

$$\mathbf{A}(z^{-1})\mathbf{y}[k]=0 \tag{4.140}$$

– ARMAX model described by

$$\mathbf{A}(z^{-1})\mathbf{y}[k]=\mathbf{B}(z^{-1})\mathbf{u}[k-nk]+\mathbf{C}(z^{-1})\mathbf{e}[k] \tag{4.141}$$

– Box-Jenkins model

$$\mathbf{y}[k]=\frac{\mathbf{B}(z^{-1})}{\mathbf{F}(z^{-1})}\mathbf{u}[k-\delta]+\frac{\mathbf{C}(z^{-1})}{\mathbf{D}(z^{-1})}e[k] \tag{4.142}$$

where

$$\mathbf{C}(z^{-1})=c_0+c_1z^{-1}+... \tag{4.143}$$

$$\mathbf{D}(z^{-1})=d_0+d_1z^{-1}+... \tag{4.144}$$

$$\mathbf{F}(z^{-1})=f_0+f_1z^{-1}+... \tag{4.145}$$

and δ is a number of delaying steps between the input and output.

4.11 Implementing Black-Box Identification Using MATLAB

One of the models listed in *System Identification Toolbox* library of MATLAB software is the model

$$\mathbf{A}(q^{-1})\mathbf{y}[k]=\frac{\mathbf{B}(q^{-1})}{\mathbf{F}(q^{-1})}\mathbf{u}[k-nk]+\frac{\mathbf{C}(q^{-1})}{\mathbf{D}(q^{-1})}\mathbf{e}[k] \tag{4.146}$$

Its structure is shown in Fig. 4.21. The models (4.132) and (4.140) – (4.142) are the special cases of (4.146).

We apply the black-box method to the ARX model and the virtual object defined by the discrete Laplace transform. Using MATLAB, the identification experiment is carried out in the following steps:

In the first step the measuring system in Simulink program is set up – Fig. 4.22. Its *Subsystem* block is shown in Fig. 4.23. The *Sign* block executes $y=sign(x)$ relation while *Band-Limited White Noise* is the white noise generator. Generation method of this signal is described in the menu under *Seed*. The block *Transfer Fcn* represents the digital transfer function of identified object.

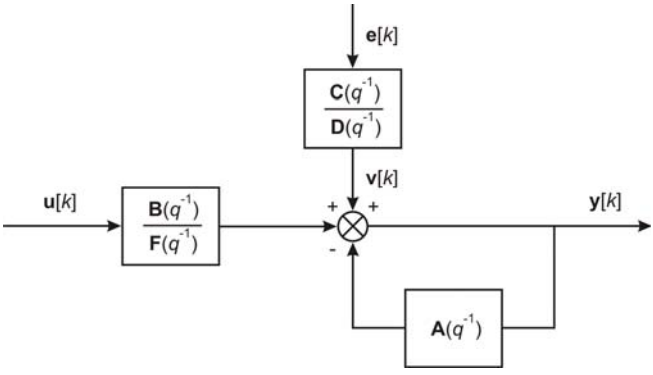

Fig. 4.21 Model structure applied in Identification Toolbox library

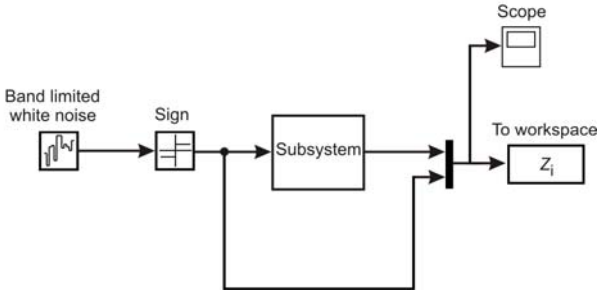

Fig. 4.22 Measuring system in Simulink program

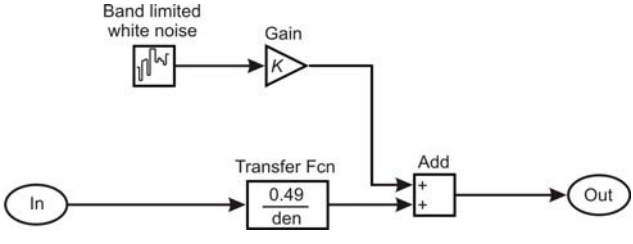

Fig. 4.23 Internal structure of *Subsystem* block

In the second step by means of *To Workspace* block, measured data are transferred and read into MATLAB working environment, and recorded as the $\mathbf{Z}_1$ matrix. The vectors $u[k]$ and $y[k]$ are used during identification process

$$\mathbf{Z_1} = \begin{bmatrix} y_1[k] & u_1[k[\\ . & . \\ y_m[k] & u_m[k] \end{bmatrix} \tag{4.147}$$

As an example, obtained through the command

>> ***idplot***($\mathbf{Z_1}$)

for *den* = [1 4.2 0.49] and *Seed* = [1 2 3 1 2], data of the first measured series of 300 samples, are shown in Fig. 4.24.

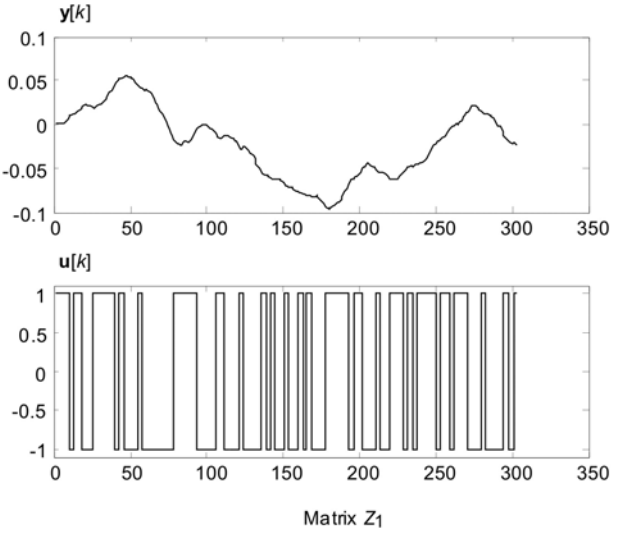

Fig. 4.24 Examples of data from second measuring series

The second series of measurements is used for a verification of the given model, and is recorded in $\mathbf{Z_2}$ matrix. However, the setup of *Seed* block must be changed before starting these measurements. Examples of measurement results $\mathbf{Z_2}$ for *Seed* = [4 3 4 1 2] are shown in Fig. 4.25.

In the third step errors generated by noise and random trends are removed. Completing this task is possible using the ***trend*** function, for which

>> $\mathbf{Z_{11}}$ = ***dtrend***($\mathbf{Z_1}$)

and

>> $\mathbf{Z_{22}}$ = ***dtrend***($\mathbf{Z_2}$)

are corresponding matrices of processed data contained in $\mathbf{Z_1}$ and $\mathbf{Z_2}$. The input and output functions, obtained through the applied ***dtrend*** function, are shown in Fig. 4.26 and Fig. 4.27.

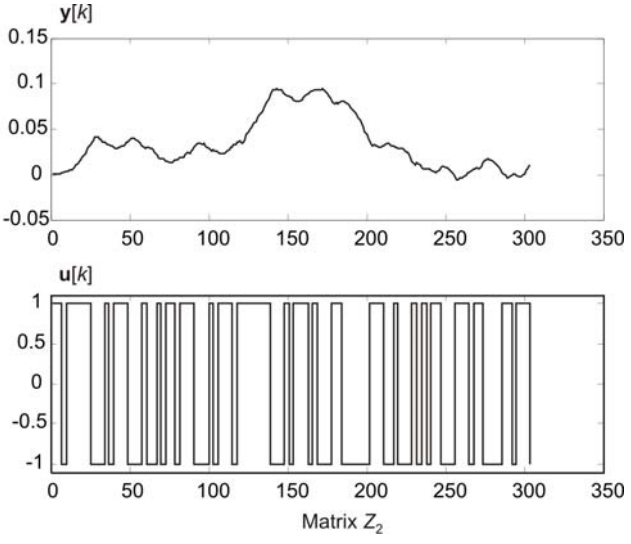

Fig. 4.25 Examples of data from second measured series

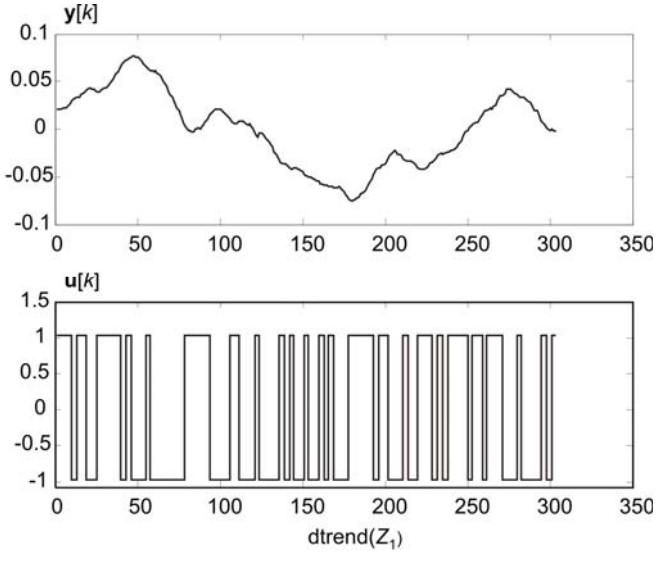

Fig. 4.26 Input and output functions obtained through ***dtrend*($\mathbf{Z_1}$)**

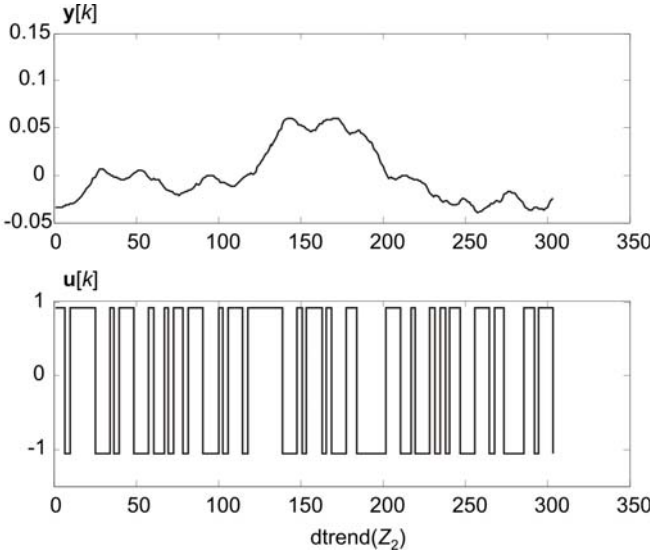

Fig. 4.27 Input and output functions obtained through ***dtrend***($\mathbf{Z}_2$)

The step four refers to the determination of model parameters. The structure of the model is determined using $\mathbf{Z}_{11}$ matrix, and parameters are estimated through the least squares estimation method. The model ARX is identified by means of the function

>> ***th*** = ***arx***($\mathbf{Z}_{11}$, $\boldsymbol{\varphi}$)

where the initial values of the vector $\boldsymbol{\varphi} = [n \;\; m \;\; \boldsymbol{\delta}]$ are fixed as [1 1 1]. The model structure and parameters, the quality coefficient applied and the number of inputs and outputs of model is displayed in the matrix ***th***. This can be achieved using the instruction

>> ***present***(***th***)

The last step refers to the model verification. A response of the model is compared with a response of the identified object. The measured data, recorded in the $\mathbf{Z}_{22}$ matrix, is used for it. The comparison is expressed through the value of *Fit* coefficient, and the comparing function is

>> ***compare***($\mathbf{Z}_{22}$, ***th***)

If the value of coefficient is not satisfying the requirements, the values of vector $\boldsymbol{\varphi}$ parameters should be changed. Examples of the model verification for $\boldsymbol{\varphi} = [1 \;\; 1 \;\; 1]$ and $\boldsymbol{\varphi} = [3 \;\; 3 \;\; 1]$ are shown in Fig. 4.28 and Fig.4.29.

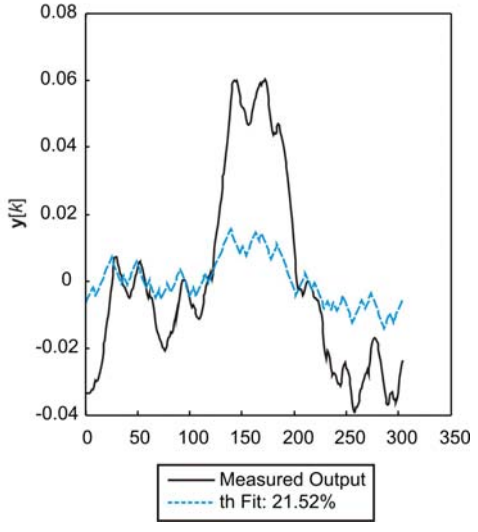

Fig. 4.28 Model [1, 1, 1] verification

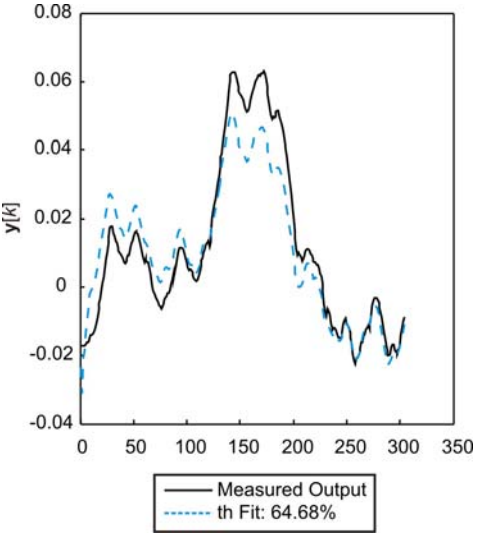

Fig. 4.29 Model [3, 3, 1] verification

In MATLAB is also available the function which allows automatically calculate *Fit* coefficient for models defined by means of vector φ. It is

>> ***arxstruc***($\mathbf{Z}_{11}$, $\mathbf{Z}_{22}$, φ_{max})

where matrix $\boldsymbol{\varphi}_{\max}$ is defined as follows

$$>> \boldsymbol{\varphi}_{\max} = \textbf{\textit{struc}}\,(1 \div n_{\max}, 1 \div m_{\max}, 1 \div \delta_{\max})$$

and

$$\boldsymbol{\varphi}_{\max} = \begin{bmatrix} 1 & 1 & 1 \\ 1 & 1 & 2 \\ 1 & 1 & 3 \\ . & . & . \\ n_{\max} & m_{\max} & \delta_{\max} \end{bmatrix} \tag{4.148}$$

4.12 Monte Carlo Method

Using the Monte Carlo method, it can be noticed that good results of parameter identification can be achieved at the relatively small amount of work required. The Monte Carlo performance will be shown using data obtained from the measuring system of Fig. 4.17 as an example. A few steps of the procedure can be distinguished.

At first, using data from measurements and our intuition, we decide and select values of the parameters' vector

$$\mathbf{W} = [a_1, a_2, ..., a_n] \tag{4.149}$$

of the assumed model

$$Y(x) = f(x, \mathbf{W}) \tag{4.150}$$

with the defined estimate-error

$$\mathbf{W\delta} = [a_1 \pm \delta_1, a_2 \pm \delta_2, ..., a_n \pm \delta_n] \tag{4.151}$$

In the second step, we select and choose a generator of pseudorandom numbers. The vector $\mathbf{W\delta}$ will be selected at random for the intervals defined by error-margins $\delta_1, \delta_2, ..., \delta_n$ (see 4.151). The user determines the number of samples. Usually, it is within the interval $K \in (10^4 - 10^6)$.

If the MathCad software is used, the function *runif* can be quoted as an example of sampling process discussed above. The function generates pseudorandom numbers of the uniform distribution. The way to use it is shown below

$$a = runif\,(K, -\delta, +\delta) \tag{4.152}$$

During the third step, the matrix $\boldsymbol{\Phi}$ is determined for the discrete values $Y(x)$. The values $Y(x)$ are calculated for parameters of the vector $\mathbf{W\delta}$ and for the $i-th$ value of the vector of successive measuring points

$$\mathbf{X} = (x_1, x_2, x_3, ..., x_m) \tag{4.153}$$

during the $k-th$ sampling

$$\mathbf{\Phi} = \begin{bmatrix} Y_{1,1} & Y_{1,2} & . & Y_{1,k} & . & Y_{1,K} \\ Y_{2,1} & Y_{2,2} & . & Y_{2,k} & . & Y_{2,K} \\ . & . & . & . & . & . \\ Y_{m,1} & Y_{m,2} & . & Y_{m,k} & . & Y_{m,K} \end{bmatrix} \tag{4.154}$$

where $k = 1, 2, ..., K$, $i = 1, 2, ..., m$, m – the number of measuring points.

During the fourth step, the matrix $\mathbf{\Delta}$ of model error values is calculated. It is determined by subtracting the vector $\mathbf{Y}$ from the consecutive columns of the matrix $\mathbf{\Phi}$, where the vector $\mathbf{Y}$ is given by

$$\mathbf{Y} = (y_1, y_2, y_3, ..., y_m)^{\mathrm{T}} \tag{4.156}$$

where the vector of measuring data $\mathbf{Y}$ corresponds to $\mathbf{X}$.

$$\mathbf{\Delta} = \begin{bmatrix} Y_{1,1} - y_1 & Y_{1,2} - y_1 & . & Y_{1,k} - y_1 & . & Y_{1,K} - y_1 \\ Y_{2,1} - y_2 & Y_{2,2} - y_2 & . & Y_{2,k} - y_2 & . & Y_{2,K} - y_2 \\ . & . & . & . & . & . \\ Y_{m,1} - y_m & Y_{m,2} - y_m & . & Y_{m,k} - y_m & . & Y_{m,K} - y_m \end{bmatrix} \tag{4.157}$$

In the next step, the least-squares method is used and the vector $\mathbf{S}$ is determined. The vector $\mathbf{S}$ is the sum of squared errors of each column of the matrix $\mathbf{\Delta}$

$$\mathbf{S} = \sum_i (\mathbf{\Delta})^2 \tag{4.158}$$

Finally, during the last step the smallest value of the vector $\mathbf{S}$ is searched for. The parameters of model related to this smallest value are assumed to be the optimal ones. The corresponding number of the sampling is also established.

References

[1] Ahlberg, J.H., Nilson, E.N., Walsh, J.L.: The Theory of Splines and Their Applications. Mathematics in Science and Engineering. Academic Press, London (1967)

[2] Allemandou, P.: Low-pass filters approximating in modulus and phase the exponential function. IEEE Transactions on Circuit Theory 13, 298–301 (1966)

[3] Arumugam, M., Ramamoorty, M.: A method of simplifying large dynamic systems. Int. J. Control 17, 1129–1135 (1973)

[4] Bockowska, M., Orlowski, M., Zuchowski, A.: O pewnej metodzie wyznaczania parametrow uproszczonych, liniowych modeli dynamiki obiektow. Pomiary Automatyka Kontrola 12, 280–282 (1994)

[5] Burden, R.L., Faires, J.D.: Numerical Analysis. PWS-KENT Publishing Company, Boston (1985)
[6] Ching-Tien, L., Yi-Shyong, C.: Successive parameter estimation of continuous dynamic systems. Int. J. Systems Science 19, 1149–1158 (1988)
[7] Director, S.W., Rother, R.A.: Introduction To Systems Theory. McGraw-Hill, New York (1972)
[8] Eykhoff, P.: System Identification. John Wiley and Sons, London (1974)
[9] Fhrumann, A.: A Polynomial Approach to Linear Algebra. Springer, New York (1996)
[10] Gawransky, W., Natke, H.G.: Order estimation of AR and APMA models. Int. J. Systems Science 19, 1143–1148 (1988)
[11] Halevi, Y.: Reduced order models with delay. Int. J. Control 64, 733–744 (1996)
[12] Hutton, M.F., Friedland, B.: Routh approximations for reducing order of linear time-invariant systems. IEEE Trans. Autom. Control 20, 329–337 (1975)
[13] Hwang, C., Lee, Y.: Multifrequency Padé Approximation Via Jordan Continued-Fraction Expansion. IEEE Trans. Autom. Control 34, 444–446 (1989)
[14] Kasprzyk, J.: Identyfikacja procesow praca zbiorowa. Wydawnictwo Politechniki Slaskiej. Gliwice (2002)
[15] Ku, Y.H.: Transient Circuit Analysis. D. Van Nostrand Co. Inc., Princeton (1961)
[16] Kubisa, S., Moskowicz, S.: A study on transitivity of Monte Carlo based evaluation of the confidence interval for a measurement result. Pomiary Kontrola Automatyka (2007)
[17] Layer, E.: Modelling of Simplified Dynamical Systems. Springer, Heidelberg (2002)
[18] Layer, E., Piwowarczyk, T.: Application of the generalized Fibonacci sequences to the simplification of mathematical models of linear dynamic systems. In: Archives of Electrical Engineering, vol. 187/188, pp. 19–30. Polish Scientific Publishers PWN, Warsaw (1999)
[19] Layer, E., Piwowarczyk, T.: Generalised Fibonacci Series in the Description of Dynamic Models. In: Systems-Analysis-Modelling-Simulation, vol. 37, pp. 57–67. Gordon & Breach Science Publisher (2000)
[20] Luke, Y.L.: The Special Functions and their Approximations. Academic Press, New York (1969)
[21] Luke, Y.L.: Mathematical Functions and their Approximations. Academic Press, New York (1975)
[22] Meier, L., Luenberger, D.G.: Approximation of linear constant systems. IEEE Trans. Autom. Control 12, 585–588 (1967)
[23] Piwowarczyk, T.: Multipower Notation of Symmetric Polynomials in Engineering Calculus. Wydawnictwo Instytutu Gospodarki Surowcami Mineralnymi i Energia PAN, Krakow (2000)
[24] Riley, K.F., Hobson, M.P., Bence, S.J.: Mathematical Methods for Physical and Engineering, 3rd edn. Cambridge University Press, Cambridge (2006)
[25] Sinha, N.K., De Bruin, H.: Near optimal control of high-order systems using low-order models. Int. J. Control 17, 257–262 (1973)
[26] Wesołowski, J.: Metoda wyznaczania klasy dynamiki obiektow liniowych. Pomiary Automatyka Robotyka, 7–8 (2004)
[27] Wesołowski, J.: Nowa metoda okreslania zastepczego rzedu modelu dynamiki. Pomiary Automatyka Kontrola, 7–8 (2004)
[28] Zuchowski, A.: O pewnej metodzie wyznaczania parametrow Strejca. Pomiary Automatyka Kontrola 2, 33–35 (1993)

[29] Zuchowski, A.: Uproszczone modele dynamiki. Wydawnictwo Uczelniane Politechniki Szczecinskiej. Szczecin (1998)
[30] Zuchowski, A.: Wyznaczenie parametrow rozszerzonego modelu Strejca w oparciu o pomiar charakterystyki skokowej. Pomiary Automatyka Kontrola 7, 6–9 (2000)
[31] Zuchowski, A.: Modele dynamiki i identyfikacja. Wydawnictwo Uczelniane Politechniki Szczecińskiej. Szczecin (2003)
[32] Zuchowski, A.: Przeglad metod wyznaczania parametrow dynamiki obiektow oscylacyjnych w oparciu o pomiar charakterystyki skokowej. Pomiary Automatyka Kontrola 9, 181–182 (2007)

Chapter 5
Mapping Error

Mapping error of models can easily be determined, if some initial information is given, like mathematical description of models, an input signal and an error criterion. Things are more complicated, when we consider object models operating in the dynamic mode. Then we deal with signals, which cannot be determined and their shapes cannot be predicted in advance.

As it is impossible to analyse the full range of all possible dynamic input signals, we have to narrow the number of signals to be considered. The immediate question is about criteria of signal selection, i.e. which signals should be used to determine mapping errors for systems with dynamic input signals of unknown both-shape and spectral distribution.

The answer lies in the concept of approaching the problem in a different way. Instead of selecting a special group of signals, we will find out the one, which will represent all signals of our interest. It is the signal generating errors of maximum value. Any other signal of any shape will generate smaller, or equal, error. This way all possible input signals to a real object will be included in this special one.

The existence and availability of signals maximizing both the integral square error and the absolute value of error are discussed, and the solutions are presented in this chapter. Constrains imposed on the input signal are also considered. These constraints refer to magnitude as well as to maximum rate of signal change. The last constraint is applied in order to match the dynamic properties of the signal to the dynamic properties of the object.

5.1 General Assumption

Let the mathematical model of a object be given by the state equation

$$\begin{aligned} \dot{x}_m(t) &= \mathbf{A}_m x(t) + \mathbf{B}_m u(t) \qquad x_m(0) = 0 \\ y_m(t) &= \mathbf{C}_m^T x(t) \end{aligned} \tag{5.1}$$

and the object, which is its reference, be given by a similar equation

$$\begin{aligned} \dot{x}_r(t) &= \mathbf{A}_r x(t) + \mathbf{B}_r u(t) \qquad x_r(0) = 0 \\ y_r(t) &= \mathbf{C}_r^T x(t) \end{aligned} \tag{5.2}$$

Let us introduce a new state equation

$$\begin{aligned} \dot{x}(t) &= \mathbf{A}x(t) + \mathbf{B}u(t) \\ y(t) &= \mathbf{C}^T x(t) \end{aligned} \tag{5.3}$$

in which

$$x(t) = \begin{bmatrix} x_r(t) \\ x_m(t) \end{bmatrix} \quad \mathbf{A} = \begin{bmatrix} \mathbf{A}_r & 0 \\ 0 & \mathbf{A}_m \end{bmatrix} \quad \mathbf{B} = \begin{bmatrix} \mathbf{B}_r \\ \mathbf{B}_m \end{bmatrix} \quad \mathbf{C} = \begin{bmatrix} \mathbf{C}_r \\ -\mathbf{C}_m \end{bmatrix} \tag{5.4}$$

where in $(5.1)-(5.4)$ $u(t)$ and $y(t)$ are the input and output respectively, $\mathbf{A}, \mathbf{B}, \mathbf{C}$ are the real matrices of corresponding dimensions.

5.2 Signals Maximizing the Integral Square Error

5.2.1 Existence and Availability of Signals with Two Constraints

Let us assume that U is the set of signals $u(t)$ piecewise continuous over the interval $[0,T]$, and the error $y(t)$ is expressed by inner product

$$I(u) = \int_0^T y^2(t)\,dt = (Ku, Ku) \qquad u \in U \tag{5.5}$$

where

$$Ku = y(t) = \int_0^t k(t-\tau)\,u(\tau)d\tau \tag{5.6}$$

and

$$k(t) = \mathbf{C}^T e^{\mathbf{A}t}\mathbf{B} \tag{5.7}$$

Let us consider the signal $h \in U$ and let the following condition be fulfilled

$$\forall 0 < b < c < T \quad \exists h \in U : supp\, h \subset [b,c] \tag{5.8}$$

and the positive square functional

$$I(h) > 0 \tag{5.9}$$

Let us define the following set U of signals with imposed constraints on the magnitude a and the rate of change ϑ

$$A = \{u(t) \in U : |u(t)| \le a,\ |\dot{u}_+(t)| \le \vartheta,\ |\dot{u}_-(t)| \le \vartheta,\ t \in [0,T]\} \tag{5.10}$$

where $\dot{u}_+(t)$ and $\dot{u}_-(t)$ are increasing and decreasing derivatives of $u(t)$ respectively.

Let $u_0(t) \in U$ fulfills the condition

$$I(u_0) = \sup\{I(u) : u \in U\} \tag{5.11}$$

then

Theorem

$$\forall t \in [0,T] \;\; |u_0(t)| = a \;\; or \;\; |\dot{u}_{0+}(t)| = \vartheta \;\; or \;\; |\dot{u}_{0-}(t)| = \vartheta \tag{5.12}$$

Proof

Suppose that (5.12) is not true. Then

$$\exists \varepsilon > 0, \quad \exists 0 < b < c < T \tag{5.13}$$

such, that

$$|u_0(t)| \le a - \varepsilon, \; |\dot{u}_{0+}(t)| \le \vartheta - \varepsilon, \; |\dot{u}_{0-}(t)| \le \vartheta - \varepsilon, \; t \in (b,c) \tag{5.14}$$

Let us choose h according to (5.8)

$$supp\, h \subset [b,c], \;\; I(h) > 0 \tag{5.15}$$

then $\exists \delta > 0$ and for small $d \in \Re$, say $d \in (-\delta, \delta)$ is

$$u_0 + dh \in A, \;\; \forall d \in (-\delta, \delta) \tag{5.16}$$

and from the optimal condition $u_0(t)$ it is evident that

$$I(u_0) \ge I(u_0 + dh) \tag{5.17}$$

hence

$$I(u_0) \ge I(u_0) + d^2 I(h) + 2d(Ku_0, Kh), \;\; d \in (-\delta, \delta) \tag{5.18}$$

and

$$0 \ge d^2 I(h) + 2d(Ku_0, Kh), \;\; d \in (-\delta, \delta) \tag{5.19}$$

However, the last inequality will never be fulfilled for $I(h) > 0$, $d \in (-\delta, \delta)$. So, from this contradiction it is obvious that $I(u_0)$ can only fulfill the condition (5.11), if the input signal $u_0(t)$ reaches one of the constraints given in (5.12).

Corollary

The proof presented above reduces shapes of the signals $u_0(t)$ to triangles or trapezoids, if constraints are imposed simultaneously on the magnitude and rate of

change. It means that the signals $u_0(t)$ can only take the form of triangles with the slope inclination $|\dot{u}_{0+}(t)| = \vartheta$ or $|\dot{u}_{0-}(t)| = \vartheta$, or of trapezoids with the slopes $|\dot{u}_{0+}(t)| = \vartheta$ $|\dot{u}_{0-}(t)| = \vartheta$ and the magnitude of a. Carrying out the proof in the identical way, it can be shown that if only one of the constraints is imposed on the signal, either on the magnitude a or on the rate of change ϑ, then the functional $I(u_0)$ reaches maximum, if the signal reaches this constraint over the interval $[0, T]$.

If the only constraint imposed on the signal $u_0(t)$ is the magnitude constraint, then it is of "bang-bang" type and it is possible to determine its switching moments. Below, we will present the analytical solution for determining such a signal.

5.2.2 Signals with Constraint on Magnitude

If the only constraint applied to the input signal is the constraint of magnitude the problem is limited to determining its switching instants only. In order to determine these switchings, let us consider the equation (5.5), which can be presented as follows

$$I(u) = (Ku, Ku) = (K^* Ku, u) \tag{5.20}$$

where the operator K^* is the conjugate of K

$$(K^* Ku, u) = \int_t^T k(\tau - t) \int_0^\tau k(\tau - \nu) u(\nu) d\nu \, d\tau \tag{5.21}$$

Let the signals $u \in U$ be limited in magnitude

$$|u(t)| \le a \qquad 0 < a \le 1 \tag{5.22}$$

From the condition of optimality (5.11), it is evident that

$$\left(\frac{\partial I(u)}{\partial u}\Big|_{u_0}, u - u_0 \right) \le 0 \tag{5.23}$$

Having computed the derivative $\frac{\partial I(u)}{\partial u}\Big|_{u_0}$, considering (5.20) and performing simple transformations (5.23) yields

$$(K^* Ku_0, u) \le (K^* Ku_0, u_0) \tag{5.24}$$

in which the right-hand side presents the maximum. Left-hand side of the formula (5.24) reaches maximum, making both sides equal, if a signal with a maximum permissible magnitude

$$|u_0(t)| = a \tag{5.25}$$

has the form

$$u(t) = u_0(t) = sign[K^* K u_0(t)] \tag{5.26}$$

After considering (5.21), we finally obtain

$$u_0(t) = a\, sign\left[\int_t^T k(\tau - t) \int_0^\tau k(\tau - v) u_0(v) dv\, d\tau\right] \tag{5.27}$$

The maximum value $\max I(u)$ generated by the signal $u_0(t)$ is

$$I(u_0) = \int_0^T |K^* K u_0(t)|\, dt = a^2 \int_0^T \left| \int_t^T k(\tau - t) \int_0^\tau k(\tau - v) u_0(v)\, dv\, d\tau \right| dt \tag{5.28}$$

5.2.3 Algorithm for Determining Signals Maximizing the Integral Square Error

From the formula (5.27), it comes that $u_0(t)$ is a signal of the "bang-bang" type, with maximum magnitude assuming the value of $a = +1$ or $a = -1$ by virtue of (5.25), and with the switching instants $t_1, t_2, ..., t_n$ corresponding to the consecutive $i = 1, 2, ..., n$ zeros of the function, occurring under the *sign* mark in the formula (5.27). In order to determine these instants, let us assume that the first switching of the signal $u_0(t)$ occurs from $+1$ to -1. It means that during the first time-interval of $0 < t \le t_1$, the signal $u_0(t) = +1$ Let us also assume that we will search for n switchings over the interval $[0, T]$. On the basis of the formula (5.27), we can write n equations with $t_1, t_2, ..., t_n$ as variables for those assumptions. It can be easily seen that the equations are described by the following relation

$$\sum_{l=i}^{n} \int_{t_l}^{t_{l+1}} k(\tau - t_i) \sum_{m=0}^{l} (-1^m) \int_{t_m}^{t_{m+1}} k(\tau - v) dv\, d\tau = 0 \qquad i = 1, 2, ..., n \tag{5.29}$$

where $t_0 = 0$, $t_{n+1} = T$, $t_{m+1} = \tau$ for $m = l$, n – number of switchings.

Solution of system equations (5.29) with respect to $t_1, t_2, ..., t_n$ gives the required switchings instants of the signal $u_0(t)$.

Between those instants, depending on the interval $t \le t_1$, $t_1 < t \le t_2, ..., t_n < t \le T$, function $K^* K u_0(t, t_1, ..., t_n)$ in (5.28) is determined by the system of $n+1$ following relations

$$K^* K u_0(t, t_1, t_2, ..., t_n) = \sum_{l=i-1}^{n} \int_{t_l}^{t_{l+1}} k(\tau - t) \sum_{m=0}^{l} (-1)^m \int_{t_m}^{t_{m+1}} k(\tau - \nu) d\nu d\tau \qquad (5.30)$$
$$i = 1, 2, ..., n+1$$

where $t_l = t$ for $l = i-1$.

The value $I(u_0)$ is determined by the sum of modules, which is determined by the formula (5.30) over all $n+1$ intervals

$$I(u_0) = \sum_{i=1}^{n+1} \left| \sum_{l=i-1}^{n} \int_{t_l}^{t_{l+1}} k(\tau - t) \sum_{m=0}^{l} (-1)^m \int_{t_m}^{t_{m+1}} k(\tau - \nu) d\nu d\tau \right| \qquad (5.31)$$

Exemplary equations for $n = 3$ switching instants in t_1, t_2 and t_3 resulting from formulae (5.29) and (5.30) are as follows:

From (5.29), we have three equations

$$\begin{aligned} &\int_{t_1}^{t_2} k(\tau - t_1) \left[\int_0^{t_1} k(\tau - v) dv - \int_{t_1}^{\tau} k(\tau - v) dv \right] d\tau \\ &+ \int_{t_2}^{t_3} k(\tau - t_1) \left[\int_0^{t_1} k(\tau - v) dv - \int_{t_1}^{t_2} k(\tau - v) dv + \int_{t_2}^{\tau} k(\tau - v) dv \right] d\tau \\ &+ \int_{t_3}^{T} k(\tau - t_1) \left[\int_0^{t_1} k(\tau - v) dv - \int_{t_1}^{t_2} k(\tau - v) dv + \int_{t_2}^{t_3} k(\tau - v) dv - \int_{t_3}^{\tau} k(\tau - v) dv \right] d\tau = 0 \end{aligned} \qquad (5.32)$$

$$\begin{aligned} &\int_{t_2}^{t_3} k(\tau - t_2) \left[\int_0^{t_1} k(\tau - v) dv - \int_{t_1}^{t_2} k(\tau - v) dv + \int_{t_2}^{\tau} k(\tau - v) dv \right] d\tau \\ &+ \int_{t_3}^{T} k(\tau - t_2) \left[\int_0^{t_1} k(\tau - v) dv - \int_{t_1}^{t_2} k(\tau - v) dv + \int_{t_2}^{t_3} k(\tau - v) dv - \int_{t_3}^{\tau} k(\tau - v) dv \right] d\tau = 0 \end{aligned} \qquad (5.33)$$

$$\int_{t_3}^{T} k(\tau - t_3) \left[\int_0^{t_1} k(\tau - v) dv - \int_{t_1}^{t_2} k(\tau - v) dv + \int_{t_2}^{t_3} k(\tau - v) dv - \int_{t_3}^{\tau} k(\tau - v) dv \right] d\tau = 0 \qquad (5.34)$$

and we have four equations resulting from (5.30):

for $0 < t \le t_1$

$$\begin{aligned} K^*Ku_0(t,t_1,t_2,t_3) &= \int_t^{t_1} k(\tau-t)\int_0^{\tau} k(\tau-v)dv\,d\tau \\ &+ \int_{t_1}^{t_2} k(\tau-t)\left[\int_0^{t_1} k(\tau-v)dv - \int_{t_1}^{\tau} k(\tau-v)dv\right]d\tau \\ &+ \int_{t_2}^{t_3} k(\tau-t)\left[\int_0^{t_1} k(\tau-v)dv - \int_{t_1}^{t_2} k(\tau-v)dv + \int_{t_2}^{\tau} k(\tau-v)dv\right]d\tau \\ &+ \int_{t_3}^{T} k(\tau-t)\left[\int_0^{t_1} k(\tau-v)dv - \int_{t_1}^{t_2} k(\tau-v)dv + \int_{t_2}^{t_3} k(\tau-v)dv - \int_{t_3}^{\tau} k(\tau-v)dv\right]d\tau \end{aligned} \tag{5.35}$$

for $t_1 < t \le t_2$

$$\begin{aligned} K^*Ku_0(t,t_1,t_2,t_3) &= \int_t^{t_2} k(\tau-t)\left[\int_0^{t_1} k(\tau-v)dv - \int_{t_1}^{\tau} k(\tau-v)dv\right]d\tau \\ &+ \int_{t_2}^{t_3} k(\tau-t)\left[\int_0^{t_1} k(\tau-v)dv - \int_{t_1}^{t_2} k(\tau-v)dv + \int_{t_2}^{\tau} k(\tau-v)dv\right]d\tau \\ &+ \int_{t_3}^{T} k(\tau-t)\left[\int_0^{t_1} k(\tau-v)dv - \int_{t_1}^{t_2} k(\tau-v)dv + \int_{t_2}^{t_3} k(\tau-v)dv - \int_{t_3}^{\tau} k(\tau-v)dv\right]d\tau \end{aligned} \tag{5.36}$$

for $t_2 < t \le t_3$

$$\begin{aligned} K^*Ku_0(t,t_1,t_2,t_3) &= \int_t^{t_3} k(\tau-t)\left[\int_0^{t_1} k(\tau-v)dv - \int_{t_1}^{t_2} k(\tau-v)dv + \int_{t_2}^{\tau} k(\tau-v)dv\right]d\tau \\ &+ \int_{t_3}^{T} k(\tau-t)\left[\int_0^{t_1} k(\tau-v)dv - \int_{t_1}^{t_2} k(\tau-v)dv + \int_{t_2}^{t_3} k(\tau-v)dv - \int_{t_3}^{\tau} k(\tau-v)dv\right]d\tau \end{aligned} \tag{5.37}$$

for $t_3 < t \le T$

$$\begin{aligned} &K^*Ku_0(t,t_1,t_2,t_3) = \\ &\int_t^{T} k(\tau-t)\left[\int_0^{t_1} k(\tau-v)dv - \int_{t_1}^{t_2} k(\tau-v)dv + \int_{t_2}^{t_3} k(\tau-v)dv - \int_{t_3}^{\tau} k(\tau-v)dv\right]d\tau \end{aligned} \tag{5.38}$$

For a higher number of switchings $t_1, t_2, ..., t_n$ we can set up a relevant system of equations in a similar way. The procedure of searching for the optimum number of n switchings starts with the assumption $i = 1$, the solution of equation (5.29) in respect of t_1, and with checking the value $I(u_0)$ (5.31) corresponding to the

obtained solution. The procedure is repeated next for $i = 2,3,....$ This way the upper value of n is not given in advance, but is obtained through the consecutive increase until the $I(u_0)$, resulting from the formula (5.31), reaches the maximum. Such a situation occurs when the value $I(u_0)$, obtained for $n+1$ switchings, is not higher than the value of the error corresponding to n switchings, and any further increase of the number of switchings cannot lift it up any more. In consequence, the search for the optimal number of switchings will end at this value of n.

5.2.4 Signals with Two Constraints

For two constraints imposed on input signal, it seems to be impossible to find out an analytical solution in respect of the shape of the $u_0(t)$, and of the formula describing the maximum value of the integral square error. Therefore, we decided to lower our requirements of a very precise solution through analytical ways, and instead of it to use modern powerful computer programs.

Good results are achieved, if heuristic techniques of computation are applied, e.g. genetic algorithms. Principles of such an approach are discussed below. Fig. 5.1 shows the diagram of computer program for determine the integral square error by means of the genetic algorithms if both constraints, of the magnitude and the rate of change, are imposed simultaneously on the input signal.

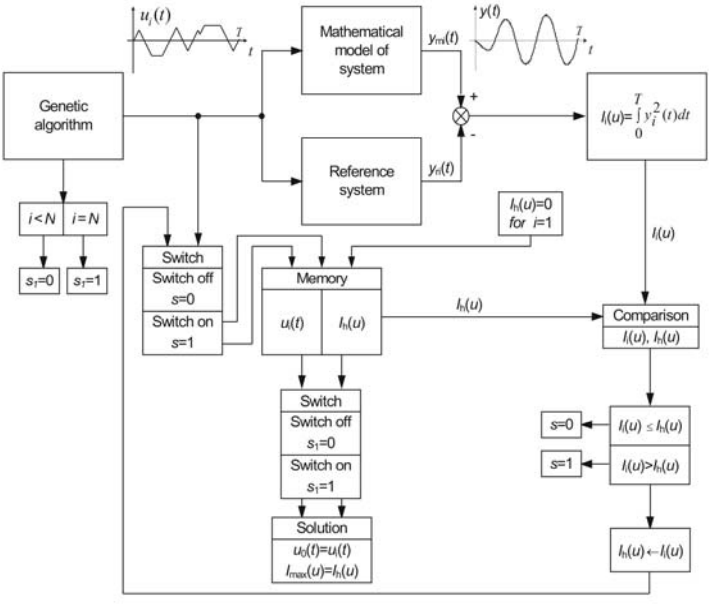

Fig. 5.1 Diagram of computer program for determining the integral square error by means of the genetic algorithm

Maximum number of iterative cycles equals

$$N = nch \cdot np \tag{5.39}$$

where nch – number of chromosomes in population, np – number of generated populations for which the stop condition is carried out.

A genetic algorithm generates, one by one, the switching vectors describing the signal $u_i(t)$, for which the error (5.5) is determined.

In every iterative cycle, the value of error $I_i(u)$ is compared with the value $I_h(u)$ stored in memory, which for $i=1$ has the initial value equal to zero. If $I_i(u) > I_h(u)$, then $I_i(u)$ is assigned to $I_h(u)$ and stored. Simultaneously with this operation, the vector of value signals $u_i(t)$ is saved in memory.

For $i = N$ the values $u_i(t)$ and $I_h(u)$ are stored in memory and are assigned to the pair of $u_0(t)$ and $I_{\max}(u)$. In this manner, the solution for $i = N$ consists of two data: the vector of data, which describes signal $u_0(t)$, and the error $I_{\max}(u) = I(u_0)$ corresponding to this value.

In order to determine signal $u_0(t)$, it is necessary to search over a set of permissible input signal $u_i(t)$.

According to specific features of genetic algorithm, determination of unknown signal $u_0(t)$ is performed in three steps:

- operation of reproduction
- operation of crossing
- operation of mutation.

Fig. 5.2 presents the flowchart of genetic algorithm.

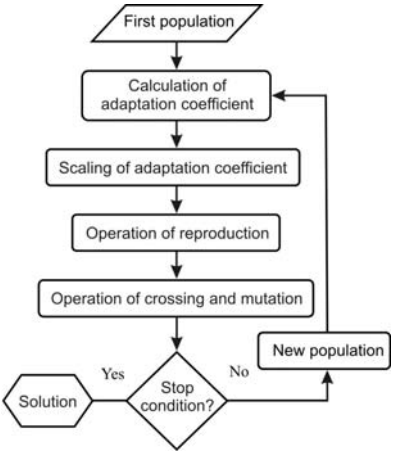

Fig. 5.2 Flowchart of genetic algorithm

During the first step, the initial population composed of an even chromosomes number is selected at random. Each chromosome consists of detectors, the number of which corresponds with the interval between switching times of $u_i(t)$. For each chromosome, the value of integral square error is determined determined – Table 5.1, and then on the basis of the obtained results and formulae (5.40) and (5.41), an adaptation coefficient is calculated.

This coefficient gives a percentage share of each chromosome in total error

$$I_{2s} = I_{21} + I_{22} + I_{23} + ... + I_{2n} \tag{5.40}$$

$$I'_{2s} = \frac{I_{2m}}{I_{2s}} \cdot 100[\%] \qquad m = 1, 2, ..., n \tag{5.41}$$

where I_{2s} is the total error, I'_{2s} gives the share in percent of individual adaptation coefficients in the total error.

The knowledge of adaptation coefficients is necessary for each chromosome in order to estimate their usefulness in population. In the case when the difference between the obtained values of adaptation coefficients is too small, it is necessary to carry out the operation of adaptation coefficient scaling. Otherwise the next steps of genetic algorithm would not give desirable effects.

Table 5.1 Chromosomes population and adaptation index for each chromosome

Chromosome	Detectors				Adaptation coefficient
	1	2	. . .	m	
p_1	t_{11}	t_{12}	. . .	t_{1m}	I_{21}
p_2	t_{21}	t_{22}	. . .	t_{2m}	I_{22}
. . .	. . .	. . .	. . .	. . .	. . .
p_n	t_{n1}	t_{n2}	. . .	t_{nm}	I_{2n}

In the next step, the operation of reproduction is carried out. On the base of the probability calculated by means of (5.40) – (5.41), the chromosomes, from the initial population, are selected at random. Depending on the value of adaptation coefficient, a particular chromosome has a larger or smaller chance to be found in the next generation. There are several ways of calculating the chances for each chromosome. The most common way is represented by the roulette wheel method. The process of random selection is carried out as many times as the number of chromosomes in the population. The results of random selection are rewritten to the new descendant population. All chromosomes have various random selection probabilities, proportional to the value of adaptation coefficient. As a result of the reproduction process, a new population composed of chromosomes $p'_1, p'_2, ..., p'_n$ is obtained.

The next step is the crossing process. Chromosomes of $p'_1, p'_2, ..., p'_n$ are joined in pairs in a random way, and for the given crossing probability P_k the number from the range [0, 1] is selected at random. If the selected number is in the range $[0, P_k]$, then the crossing process is performed. Otherwise the equivalent detectors of joined chromosomes are not crossed. The crossing probability P_k is usually established at a high level, which is about 0.9.

The crossing process is carried out according to the following formulae:

1. In the case of crossing detectors t_{11} from the first chromosome, and t_{21} from the second chromosome, we have

$$\begin{aligned} t'_{11} &= (1-\alpha)t_{11} + t_{21} \\ t'_{21} &= \alpha t_{11} + (1-\alpha)t_{21} \end{aligned} \tag{5.42}$$

where t'_{11} is a descendant detector of the first chromosome, and t'_{21} is a descendant detector of the second chromosome.

The coefficient α is selected according to the following formulae

$$\begin{aligned} \alpha_1 &= \frac{-t_{11}}{t_{21}-t_{11}} \qquad & \alpha_2 &= \frac{t_{12}-t_{11}}{t_{21}-t_{11}} \\ \alpha_3 &= \frac{-t_{21}}{t_{11}-t_{21}} \qquad & \alpha_4 &= \frac{t_{22}-t_{21}}{t_{11}-t_{21}} \end{aligned} \tag{5.43}$$

where α_1 and α_2 present the minimum and maximum limit of the α coefficient changeability for the detector from the first chromosome, while α_3 and α_4 present minimum and maximum limit of α coefficient changeability for the detector from the second chromosome.

The changeability range of α is contained in the range between zero and the third value of $\alpha_{\max-1}$ coefficient (5.44) minus $\alpha_{\max-1}$ multiplied by the changeability step of t from interval [0 ,T].

Then the value of α is selected at random from the above range, and is substituted into (5.42).

2. In the case of crossing of the detectors t_{1m} from the first chromosome and t_{2m} from the second chromosome, we have:

$$\begin{aligned} t'_{1m} &= (1-\alpha)t_{1m} + t_{2m} \\ t'_{1m} &= \alpha t_{1m} + (1-\alpha)t_{2m} \end{aligned} \tag{5.44}$$

where $m = 2, 3, ..., n$ and t'_{1m} is the first chromosome descendant m detector, and t'_{2m} is the second chromosome descendant m detector.

The coefficient α is selected according to the following formulae

$$\alpha_1 = \frac{t_{1m-1} - t_{1m}}{t_{2m} - t_{1m}} \qquad \alpha_2 = \frac{t_{1m+1} - t_{1m}}{t_{2m} - t_{1m}}$$
$$\alpha_3 = \frac{t_{2m-1} - t_{2m}}{t_{1m} - t_{2m}} \qquad \alpha_4 = \frac{t_{2m+1} - t_{2m}}{t_{1m} - t_{2m}} \tag{5.45}$$

The operation of crossing is presented in Fig. 5.3.

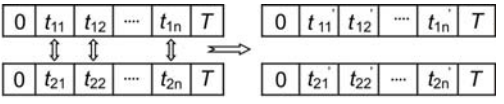

Fig. 5.3 Operation of crossing

The crossing procedure described by formulae $(5.43)-(5.46)$ assures that in the descendant chromosomes the subsequent detectors will have the value larger than the value of the detectors situated immediately before them. This requirement must be met, because individual detectors included in the chromosome contain the interval of switching times of the signal $u_i(t)$.

The operation of mutation is the last step of the genetic algorithm. In the case of each detector included in the descendant chromosomes, we ask whether the mutation operation will be carried out or not. This process usually is carried out at small probability $(P_m < 0.01)$. Mutation is a sort of supplement to the operation of crossing. There are many varieties of mutation, and the choice of relevant mutation depends on the algorithm application. The linear mutation described by formula (5.46) is often applied

$$d''_{1m} = (d'_{1m+1} - d'_{1m-1})\alpha + d'_{1m-1}, \qquad \alpha \in [0, 1], \qquad m = 1, 2, ..., n \tag{5.46}$$

The operation of mutation is presented in Fig. 5.4.

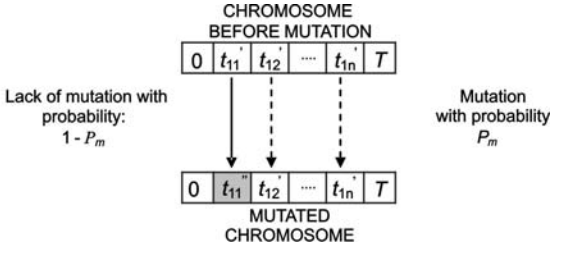

Fig. 5.4 Operation of mutation

Completing the operation of mutation, the genetic algorithm process starts again. It runs in a loop as shown in Fig. 5.2. The number of populations should be as large as possible. However, it must be noted that increasing the number of populations makes the time of genetic algorithm calculations longer. The time can be reduced significantly, if a stop condition is applied. This condition stops the algorithm if the value of $I_h(u)$ stored in memory does not change.

5.2.5 Estimation of the Maximum Value of Integral Square Error

Let us assume that the upper limit of the integral in (5.5) tends to infinity $T \to \infty$ and let the error $I(u)$ be presented as follows

$$I(u) = \int_0^{+\infty} y^2(t)dt = \int_0^{\infty} \frac{1}{j2\pi} \int_{-j\infty}^{j\infty} Y(j\omega)\, e^{j\omega t} dj\omega\, y(t)dt \tag{5.47}$$

The relation between $y(t)$ and $Y(j\omega)$ is expressed by means of Fourier transform

$$y(t) = \frac{1}{j2\pi} \int_{-j\infty}^{j\infty} Y(j\omega)\, e^{j\omega t} dj\omega \tag{5.48}$$

Changing the order of integration in (5.47), we have

$$I(u) = \frac{1}{j2\pi} \int_{-j\omega}^{+j\omega} Y(j\omega) \int_0^{+\infty} y(t)\, e^{j\omega t} dt\, dj\omega \tag{5.49}$$

hence

$$I(u) = \frac{1}{j2\pi} \int_{-j\omega}^{+j\omega} Y(j\omega) Y(-j\omega)\, dj\omega \tag{5.50}$$

and

$$I(u) = \frac{1}{2\pi} \int_{-\infty}^{+\infty} \left|Y(j\omega)\right|^2 d\omega \tag{5.51}$$

Taking into account that

$$Y(j\omega) = K(j\omega)\, X(j\omega) \tag{5.52}$$

we finally have

$$I(u) = \frac{1}{2\pi} \int_{-\infty}^{+\infty} \left|G(j\omega)\right|^2 \left|X(j\omega)\right|^2 d\omega \le \sup_{\omega} \left|G(j\omega)\right|^2 \frac{1}{2\pi} \int_{-\infty}^{+\infty} \left|X(j\omega)\right|^2 d\omega \tag{5.53}$$

and

$$\sup_{\omega} \left|G(j\omega)\right|^2 \frac{1}{2\pi} \int_{-\infty}^{+\infty} \left|X(j\omega)\right|^2 d\omega = E \sup_{\omega} \left|G(j\omega)\right|^2 \tag{5.54}$$

where E is the energy of the input signal.

Note that the estimation based on Eq. 5.54 may be many times greater than the value precisely calculated with the use of left hand side of Eq. 5.53.

5.3 Signals Maximizing the Absolute Value of Error

5.3.1 Signals with Constraint on Magnitude

In order to determine the signal maximizing the value of absolute error, let us take a convolution integral (5.55) into consideration

$$y(t) = \int_0^t k(t-\tau)u(\tau)d\tau \qquad t \in [0, T] \tag{5.55}$$

It is obvious that the maximum $|y(t)|$ occurs for $t = T$

$$\max|y(t)| = y(T) = \int_0^T k(T-\tau)u(\tau)\,d\tau \tag{5.56}$$

if

$$u(\tau) = u_0(\tau) = a \cdot sign[k(T-\tau)] \tag{5.57}$$

where a is the magnitude of $u(\tau)$.

Replacing τ by t in (5.56), we can write

$$|y(t)| = y(T) = \int_0^T k(T-t)u(t)\,dt \tag{5.58}$$

and $u_0(t)$ maximizing (5.58) has now the form

$$u_0(t) = a \cdot sign[k(T-t)] \tag{5.59}$$

Substituting (5.59) into (5.58) gives finally

$$\max|y(t)| = y(T) = a \cdot \int_0^T |k(T-t)|dt = a \cdot \int_0^T |k(t)|dt \tag{5.60}$$

which is not difficult to compute.

5.3.2 Shape of Signals with Two Constraints

Let us present the signal $u(t)$ by means of the integral

$$u(t) = \int_0^t \varphi(\tau)d\tau \tag{5.61}$$

and the error (5.58) in the following form

$$y(T) = \int_0^T k(T-t) \int_0^t \varphi(\tau) d\tau \, dt \tag{5.62}$$

The constraints (5.10), related to $u(t)$ for the function $\varphi(\tau)$, are as follows

$$\left| \int_0^t \varphi(\tau) d\tau \right| = |u(t)| \le a \tag{5.63}$$

and

$$|\varphi(t)| = |\dot{u}(t)| \le \vartheta \tag{5.64}$$

Changing the integration order in (5.62), we have

$$y(T) = \int_0^T \varphi(\tau) \int_\tau^T k(T-t) \, dt \, d\tau \tag{5.65}$$

and after replacing τ for t, we get finally

$$y(T) = \int_0^T \varphi(t) \int_t^T k(T-\tau) \, d\tau \, dt \tag{5.66}$$

From (5.66), it is evident that $\varphi(t)$, which maximizes $y(T)$, has the maximum magnitude $\varphi(t) = \pm\vartheta$ by virtue of the formula (5.64) if

$$\varphi(t) = sign \int_t^T k(T-\tau) d\tau \tag{5.67}$$

and $\varphi(t) = 0$, in such subintervals, for which the resulting form (5.67) between the switching moments is

$$\left| \int_0^t \varphi(\tau) d\tau \right| > a \tag{5.68}$$

Using the equations (5.61) – (5.66), we can determine signal $u(t) = u_0(t)$ in the following cases

First case

If $\left| \int_0^t f_0(\tau) d\tau \ne a \right|$ for δ varying in the intervals $[0, +\vartheta]$ and $[0, -\vartheta]$, (Fig 5.5 and Fig 5.6), where $f_0(t) = \pm\vartheta$ for $\int_t^T k(T-\tau) d\tau > +\delta$ and $\int_t^T k(T-\tau) d\tau < -\delta$ respectively, than the signal $u_0(t)$ is determined in three following steps, according to Eqs. (5.67) – (5.75).

During the first step, the "bang-bang" functions $f_1(t)$ of the magnitude $\pm\vartheta$ are determined with switching moments resulting from (5.67) – Fig.5.7

$$\begin{aligned} f_1(t) &= +\vartheta \quad \text{if} \quad \varphi(t) > 0 \\ f_1(t) &= -\vartheta \quad \text{if} \quad \varphi(t) < 0 \end{aligned} \tag{5.69}$$

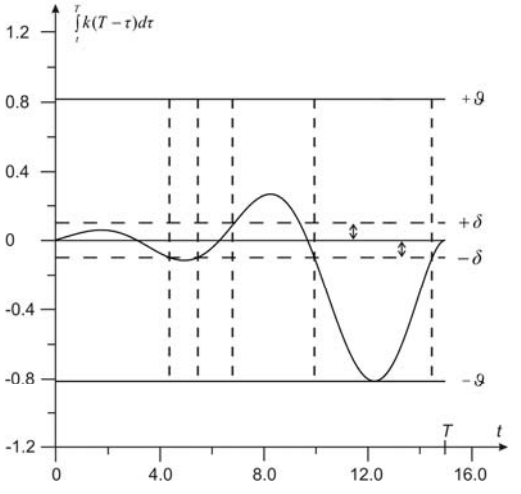

Fig. 5.5 Exemplary function $\int_t^T k(T-\tau)d\tau$

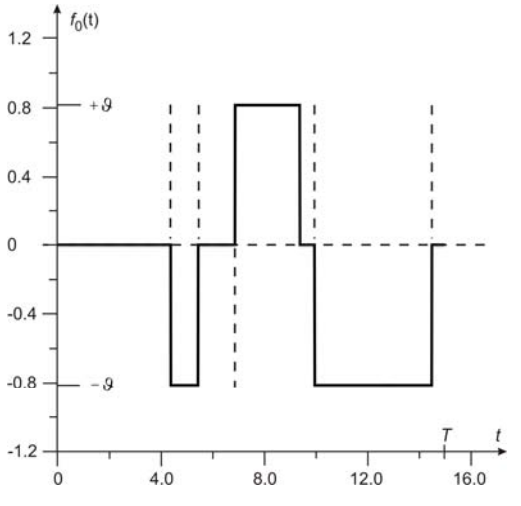

Fig. 5.6 Constraints resulting from $\int_t^T k(T-\tau)d\tau > +\delta$ and $\int_t^T k(T-\tau)d\tau < -\delta$

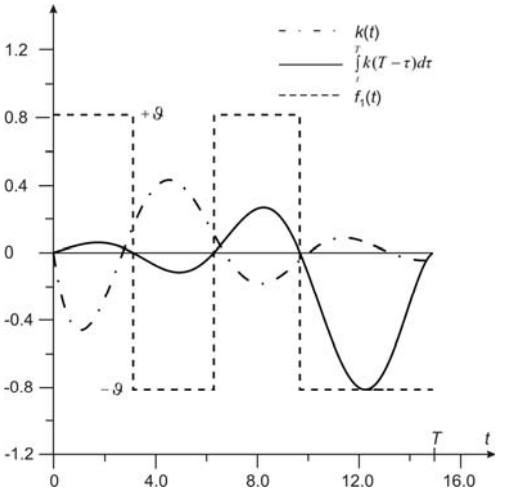

Fig. 5.7 Exemplary functions $k(t)$, $\int_t^T k(T-\tau)d\tau$ and $f_1(t)$

In the second step, we obtain the function $f_2(t)$ by integrating $f_1(t)$ – Fig. 5.8.

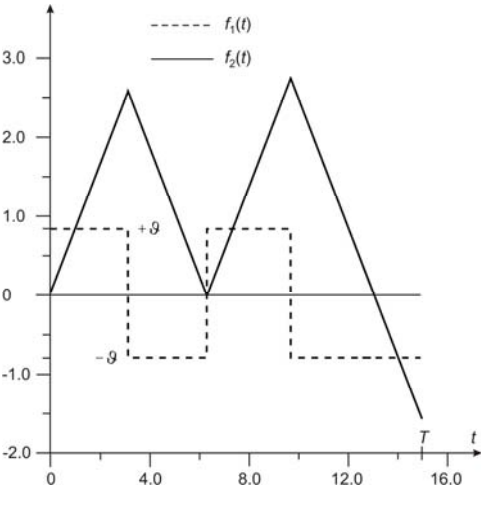

Fig. 5.8 Functions $f_1(t)$ and $f_2(t) = \int_0^t f_1(\tau)d\tau$

Function $f_2(t)$ in particular switching intervals $t_1, t_2, ..., t_n$ of $f_1(t)$ is given by the following relations

for $t \le t_1, \quad n = 1$

$$f_2(t) = \vartheta \cdot t \tag{5.70}$$

for $t_1 < t \le t_2, \quad n = 2$

$$f_2(t) = \vartheta \cdot t_1 - \vartheta \cdot (t - t_1) \tag{5.71}$$

for $t_i < t \le t_{i+1}, \quad i = 2,3,...,n, \quad t_{n+1} = T, \quad n -$ number of switchings

$$f_2(t) = \vartheta \cdot t_1 + \vartheta \cdot \sum_{j=2}^{i} (-1)^{j-1} (t_j - t_{j-1}) + (-1)^i \cdot \vartheta \cdot (t - t_i) \tag{5.72}$$

In the last step, we determine the function $f_3(t)$ on the basis of $f_2(t)$. Relation is as follows

$$\begin{aligned} f_3(t) &= \pm\vartheta \qquad \text{if} \quad |f_2(t)| \le a \\ f_3(t) &= 0 \qquad\quad \text{if} \quad |f_2(t)| > a \end{aligned} \tag{5.73}$$

Finally, we obtain the signal $u(t) = u_0(t)$ through integration of $f_3(t)$, and this is the aim. The operation is shown in Fig. 5.9.

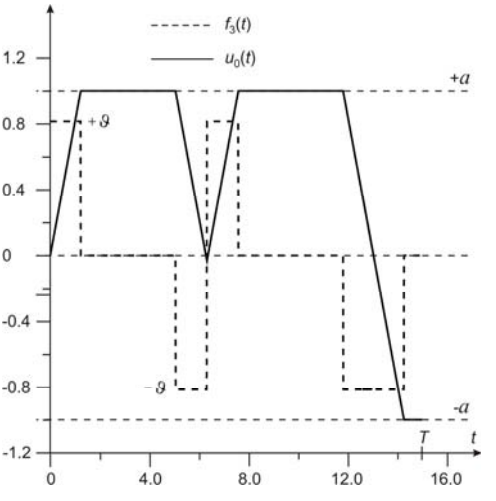

Fig. 5.9 Function $f_3(t)$ and signal $u_0(t) = \int_0^t f_3(\tau) d\tau$

During the intervals in which $f_3(t) = \pm\vartheta$, the signal shape is triangular, with the slope of $\pm\vartheta$. In the intervals when $f_3(t) = 0$, the signal is a constant of the magnitude $\pm a$.

For n switching moments of $f_3(t)$ the value of error is described by the following equations:

for $n = 1$

$$y(T) = \frac{h_1}{t_1}\int_0^{t_1} k(T-\tau)\,\tau\,d\tau + \frac{h_T - h_1}{T - t_1}\int_{t_1}^{T} k(T-\tau)\,(\tau - t_1)\,d\tau + h_1\int_{t_1}^{T} k(T-\tau)\,d\tau \tag{5.74}$$

for $n \geq 2$

$$y(T) = \frac{h_1}{t_1}\int_0^{t_1} k(T-\tau)\,\tau\,d\tau + \sum_{i=2}^{n}\left[\frac{h_i - h_{i-1}}{t_i - t_{i-1}}\int_{t_{j-1}}^{t_j} k(T-\tau)(\tau - t_{i-1})\,d\tau + h_{i-1}\int_{t_{j-1}}^{t_j} k(T-\tau)\,d\tau\right] + \frac{h_T - h_n}{T - t_n}\int_{t_n}^{T} k(T-\tau)\,(\tau - t_n)\,d\tau + h_n\int_{t_n}^{T} k(T-\tau)\,d\tau \tag{5.75}$$

where $h_i = u_0(t_i)$, $h_T = u_0(T)$.

Fig. 5.10 presents the signal $u_0(t)$ and the error $y(t)$ corresponding to it.

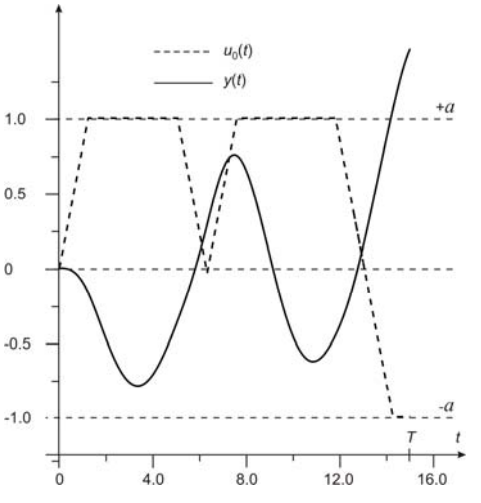

Fig. 5.10 Signal $u_0(t)$ and error $y(t)$

Second case

If $\vartheta \cdot T \le a$ then the signal $u_0(t)$ is given directly by

$$u_0(t) = \vartheta \cdot \int_0^t sign[k(T-\tau)]\,d\tau \tag{5.76}$$

and the error equals

$$y(T) = \int_0^T k(t-\tau)u_0(\tau)d\tau \tag{5.77}$$

Fig. 5.11 presents the signal $u_0(t)$ and error $y(t)$ corresponding to it.

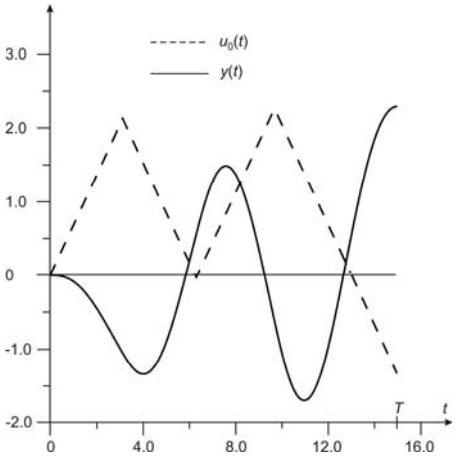

Fig. 5.11 Signal $u_0(t)$ and error $y(t)$

Third case

If $\vartheta \cdot T > a$ then the signal $u_0(t)$ is determined indirectly by means of the functions $f_4(t) - f_6(t)$

$$f_4(t) = \frac{k(t)\cdot\vartheta}{2a} \tag{5.78}$$

$$f_5(t) = a \cdot sign[f_4(T-t)] \tag{5.79}$$

$$\begin{aligned} f_6(t) &= f_5(t) && \text{if} \quad t < \frac{2a}{\vartheta} \\ f_6(t) &= f_5(t) - f_5\left(t - \frac{2a}{\vartheta}\right) && \text{if} \quad \frac{2a}{\vartheta} < t < T \end{aligned} \tag{5.80}$$

The functions $f_4(t)$ and $f_5(t)$ are shown in Fig. 5.12, while $f_6(t)$ and the signal $u_0(t) = \int_0^t f_6(\tau)\, d\tau$ in Fig. 5.13.

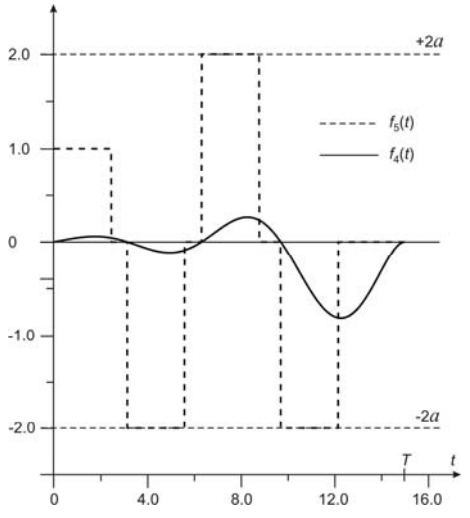

Fig. 5.12 Functions $f_4(t)$ and $f_5(t)$

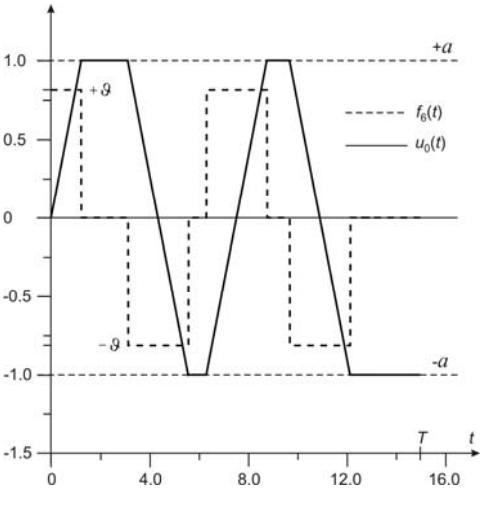

Fig. 5.13 Function $f_6(t)$ and signal $u_0(t)$

Fig. 5.14 shows the signal $u_0(t)$ and the error $y(t)$ corresponding to it.

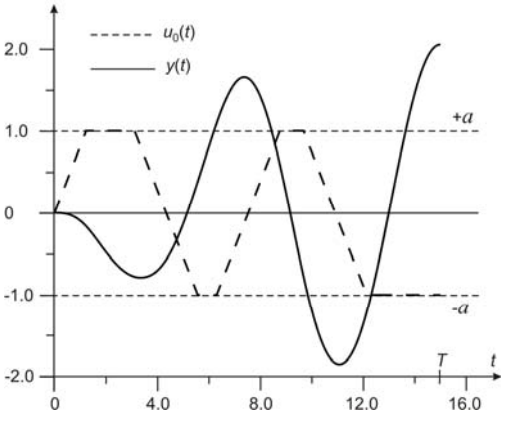

Fig. 5.14 Signal $u_0(t)$ and error $y(t)$

5.4 Constraints of Signals

Mapping errors of models are determined using precisely defined input signals. In our case, they are the signals maximizing the error and selected by a special criterion. For obvious reasons, amplitudes of such signals must always be limited. Signals limited in amplitude only of "bang-bang" type, may generate mapping errors of considerably high values, even in the situation when models are almost alike. This is caused by the particular dynamics of the "bang-bang" signals, which have derivatives of infinitely high values on the instants of switching, while outside these instants the values are constant. Such a dynamics of signals does not match the dynamics of physically existing systems, since the latter can only transmit signals with limited value of rate of change. Therefore apart from limiting the amplitude, we impose an additional constraint originating from the dynamic properties of the system under modelling.

The constraint can be determined in the time or frequency domains. If we are to consider it in the time domain, it can be assumed that the constraint refers to the maximum rate of change ϑ of the input signal. Namely, this rate is to be smaller or equal to the maximum rate of the step response of the modelled system.

$$\vartheta = \max\left|\dot{u}(t)\right| \le \max\left|\dot{h}(t)\right| = \max\left|k(t)\right| \tag{5.81}$$

where $h(t)$ and $k(t)$ denote the step and impulse responses of the system, respectively.

In the frequency domain, it is a transfer band of the system under modelling, which imposes the constrain of ϑ. Assuming the maximum harmonic ω_m of the transfer band is not distorted, we get

$$\vartheta \le \max \left| a \cdot \frac{d \sin(\omega_m t)}{dt} \right| = a \cdot \omega_m \tag{5.82}$$

However, the assessment of ω_m value is quite arbitrary very often.

References

[1] Birch, B.J., Jackson, R.: The behaviour of linear systems with inputs satisfying certain bounding conditions. J. Electronics and Control 6, 366–375 (1959)

[2] Fuksa, S., Byrski, W.: Problem optymalizacji pewnego typu funkcjonalow kwadratowych na zbiorach wypuklych. In: Prace VIII Krajowej Konferencji Automatyki, Szczecin, pp. 62–64 (1980)

[3] Goldberg, D.E.: Genetic Algorithms in Search, Optimization, and Machine Learning. Addison-Wesley Publishing Company, USA (1989)

[4] Ja, K.I.: O nakoplenii vozmuscenij v linejnych sistemach. Teoria funkcji, funkcjonalnyj analiz i ich prilozenija 3 (1967)

[5] Ja, K.I., Ulanov, G.M.: Ob efektivnom obobscenii teorii nakoplenia otklonenij. Prikladnaja matematika i mechnaika 37 (1973)

[6] Layer, E.: Theoretical foundations of the calibration process of measuring systems in the aspect of dynamic errors. In: Proc. IMEKO Symposium on Computerized Measurement, Dubrovnik, pp. 113–116 (1981)

[7] Layer, E.: Basic problems of the calibration process and of the establishing a hierarchy of accuracy for dynamic measuring systems. In: Proc. IMEKO 9-th World Congress, Berlin (West), vol. 5(3), pp. 269–277 (1982)

[8] Layer, E.: Theoretical Principles for Establishing a Hierarchy of Dynamic Accuracy with the Integral-Square-Error as an Example. IEEE Trans. Instrumentation and Measurement 46, 1178–1182 (1997)

[9] Layer, E.: Mapping Error of Simplified Dynamic Models in Electrical Metrology. In: Proc. 16-th IEEE Instrumentation and Measurement Technology Conference, Venice, vol. 3, pp. 1704–1709 (1999)

[10] Layer, E.: Mapping Error of Linear Dynamic Systems Caused by Reduced-Order Model. IEEE Trans. Instrumentation and Measurement 50, 792–800 (2001)

[11] Layer, E.: Przestrzen rozwiazan sygnalow maksymalizujących kwadratowy wskaznik jakosci. In: Mat. Konf. Modelowanie i symulacja systemow pomiarowych, pp. 25–28. AGH, Krynica (2001)

[12] Layer, E.: Modelling of Simplified Dynamical Systems. Springer, Heidelberg (2002)

[13] Layer, E.: Shapes of Input Signals for Calibration of Measuring Systems Intended for the Measurement of Dynamic Signals. In: Proc. IMEKO TC7 Symp., Cracow, Poland, pp. 146–149 (2002)

[14] Layer, E.: Non-standard input signals for the calibration and optimisation of the measuring systems. Measurement 34(2), 179–186 (2003)

[15] Layer, E., Gawedzki, W.: Dynamics of Measurement Systems. Investigation and Estimation. Polish Scientific Publisher, Warsaw (1991)

[16] Layer, E., Gawedzki, W.: Dynamika aparatury pomiarowej. Badania i ocena. PWN, Warszawa (1991)

[17] Layer, E., Tomczyk, K.: Shapes and algorithm for determining signals maximising the absolute value of error. In: Proc. V MATHMOD 2006 Conference, Vienna, Austria Mathematical and Computer Modelling of Dynamical Systems Publication, Abstract vol. 1, p. 37, full paper CD-R (2006)
[18] Ljubojevicz, M.: Suboptimal input signal for linear system identification. Int. J. Control 17, 659–669 (1973)
[19] Rutland, N.K.: The Principle of Matching: Practical Conditions for Systems with Inputs Restricted in Magnitude and Rate of Change. IEEE Trans. Autom. Control 39, 550–553 (1994)
[20] Tomczyk, K.: Optymalizacja parametrów wybranych typow przetwornikow drgan. In: Mat. Konferencyjne XXXVI Miedzyuczelnianej Konferencji Metrologow MKM 2004, Ustron 2004, Prace Komisji Metrologii PAN, Katowice, pp. 209–218 (2004)
[21] Tomczyk, K.: Optymalizacja parametrow matematycznych modeli wzorcow ze wzgledu na transformacje niezniekształcajaca. In: Mat. V Sympozjum. nt. Pomiarow Dynamicznych, Szczyrk, Prace Komisji Metrologii PAN, Katowice, pp. 119–128 (2005)
[22] Tomczyk, K.: Zastosowanie algorytmow genetycznych do wzorcowania dynamicznego aparatury pomiarowej. In: Mat. V Sympozjum. nt. Pomiarow Dynamicznych, Szczyrk, Prace Komisji Metrologii PAN, Katowice, pp. 129–139 (2005)
[23] Tomczyk, K.: Application of genetic algorithm to measurement system calibration intended for dynamic measurement. Metrology and Measurement Systems XIII(1), 193–203 (2006)
[24] Tomczyk, K.: Computer aided system for determining maximum dynamic errors of chosen measuring instruments. PhD thesis, Cracow University of Technology, Cracow (2006)
[25] Tomczyk, K., Sieja, M.: Acceleration transducers calibration based on maximum dynamic error, pp. 27–49. Czasopismo Techniczne, Politechnika Krakowska (2006)
[26] Zakian, V.: Critical systems and tolerable inputs. Int. J. Control 49, 1285–1289 (1989)
[27] Zakian, V.: Perspectives of the principle of matching and the method of inequalities. Int. J. Control 65, 147–175 (1996)

Index