



Systemics

Lecture Notes

Jürgen Pannek

June 25, 2024



Jürgen Pannek
Institute for Intermodal Transport and Logistic Systems
Hermann-Blenck-Str. 42
38519 Braunschweig



FOREWORD

This script originates from a correspondent lecture *Systemics* held during the summer term 2024 at the Technical University of Braunschweig. To structure the lecture and support my students in their learning process, I prepared these lecture notes. Within the lecture, I adapt to the students, their background and wishes, and I will integrate remarks and corrections throughout the summer term.

The aim of the module is to provide participating students with knowledge of terms of system theory and control engineering. Moreover, students shall have knowledge of terms for systems and be enabled to understand principles of system description, modeling and identification. After successfully completing the module, students shall additionally be able to apply the discussed methods and be able to assess results.

The central aims of the lecture are the introduction of modeling and system identification techniques for (dynamical) systems. In particular, we focus on the time domain and model system using differential equation systems to design

- Deterministic Processes as well as
- Stochastic Processes.

To deal with these kind of systems properly, we give a short introduction/repetition to differential equations. Based on these basic models, we then identify „the real“ system, i.e. we fit data to model. To this end, we introduce basic stochastic definitions and discuss

- Least Square Estimation and
- Kalman Filtering.

At the end of the lecture, students should understand the concepts, know basic formulas, be able to comprehend and interpret input and output of the methods and to make a suitable choice between the presented methods.

The module itself is accredited with 5 credits.

Literature for further reading

■ Background

- AULBACH, B.: *Gewöhnliche Differenzialgleichungen*. Spektrum Akademischer Verlag, 2010
- KHALIL, H.K.: *Nonlinear Systems*. Prentice Hall PTR, 2002
- SONTAG, E.D.: *Mathematical Control Theory: Deterministic Finite Dimensional Systems*. Springer, 1998

■ Growth processes

- MURRAY, J.D.: *Mathematical biology: II: Spatial models and biomedical applications*. Springer, 2003

■ Mechanical processes

- PILA, A.W.: *Introduction to Lagrangian dynamics*. Springer, 2019

■ Electrical processes

- BORUTZKY, W.: *Bond graph methodology*. Springer, 2009

■ Financial processes

- HIGHAM, D.J.: *An introduction to financial option valuation: mathematics, stochastic and computation*. Cambridge University Press, 2004
- KLOEDEN, P.E. ; PLATEN, E. ; SCHURZ, H.: *Numerical solution of SDE through computer experiments*. Springer, 2012

■ Identification

- LJUNG, L.: *System Identification: Theory for the User*. Pearson Education, 1998
- SCHOUKENS, J. ; PINTELON, R. ; ROLAIN, Y.: *Mastering System Identification in 100 Exercises*. John Wiley & Sons, 2012
- SCHOUKENS, J.: *System Identification*. Vrije Universiteit Brussel, 2013

Contents

Contents	iv
List of tables	v
List of figures	viii
List of definitions and theorems	xi
1 Introduction	1
1.1 What is a “model”?	1
1.2 What is “identification”?	5
1.3 Stochastic parameters	6
1.4 Models	11
I Modeling	15
2 Growth processes	17
2.1 Growth dynamics for one object type	17
2.1.1 From difference to differential equation	17
2.1.2 Simple growth model	18
2.1.3 Logistic growth model	21
2.2 Lotka-Volterra model	25
3 Mechanical processes	31
3.1 d’Alembert Principle	31
3.1.1 Translational elements	32
3.1.2 Rotational elements	34
3.1.3 Combining elements	37
3.2 Lagrangian formalism	39
4 Electrical processes	47

4.1	Network elements	47
4.2	Bond graph junctions	50
4.3	Bond graph modeling	53
5	Stochastic processes	61
5.1	Options	61
5.2	Monte–Carlo method	65
II	Identification	69
6	Structure of the identification process	71
6.1	Basic design of estimators	71
6.2	Properties of estimators	75
6.3	Practical differences of estimators	81
7	Least square estimation	85
7.1	Problem definition	85
7.2	Linear least square	87
7.3	Properties of the linear least square estimator	93
7.4	Weighted least square estimator	95
7.5	Properties of the weighted linear least square estimator	98
8	Kalman filtering	101
8.1	Recursive identification	101
8.2	Filter problem and assumptions	103
8.3	Propagation of mean and covariance	105
	Bibliography	113

List of Tables

1.1	Division lines for dynamic systems	4
1.2	Advantages and disadvantages of differential equations	12
1.3	Advantages and disadvantages of stochastic processes	14
2.1	Advantages and disadvantages of logistics growth model	24
2.2	Advantages and disadvantages of Lotka-Volterra	30
3.1	Denomination for technical elements and models	31
3.2	Denomination for translational models	32
3.3	Denomination for rotational models	34
3.4	Advantages and disadvantages of d'Alembert's method	39
3.5	Advantages and disadvantages of the Lagrangian/Hamiltonian approach	45
4.1	Direct analogy for power	50
4.2	C and I energy storages	56
4.3	Advantages and disadvantages of bond graphs	59
5.1	Advantages and disadvantages of Monte-Carlo	68
6.1	Advantages and disadvantages of estimation	84
7.1	Advantages and disadvantages of least squares	95
7.2	Advantages and disadvantages of weighted least squares	99
8.1	Advantages and disadvantages of Kalman filtering	112

List of Figures

1.1	Model and environment	2
1.2	Dimensions of model characteristics	3
1.3	General structure of a dynamic system	4
1.4	Sketch of the Gaussian distribution	10
1.5	Sketch of a dynamic flow and a trajectory	13
1.6	Different paths of a Wiener process	14
2.1	Growth of the world population and solution of (2.3) for identified parameters . .	20
2.2	Growth of the European population and solution of (2.3) for identified parameters	20
2.3	Solutions of the logistics equation (2.4)	26
2.4	Solutions for the Lotka-Volterra model (2.7) with $a = c = 1$	27
2.5	Solutions for the predator-prey model (2.9) with $a = c = 1$ and $\beta = 0.5$	29
3.1	Symbol for a mass element	32
3.2	Symbols for a spring element	33
3.3	Symbol for a damper element	34
3.4	Schematic illustration of torque	35
3.5	Schematic illustration of rotational mass element	36
3.6	Symbols for a rotational spring element	36
3.7	Symbol for a rotational damper element	37
3.8	Schematic drawing of a pendulum	38
4.1	Graph of network from Task 4.2	48
4.2	Example of a bond graph	49
4.3	Convention for annotation of power bonds	49
4.4	Convention for annotation of power bonds	51
4.5	Example of 0- and 1-junctions	52
4.6	Representation of a two-port transformer	54
4.7	Representation of a two-port gyrator	55
4.8	Two-port gyrator according to Task 4.17	55
4.9	Schematic of a DC machine moving a (hoisting) drum roll	58

4.10	Bond graph of DC driven drum roll	59
5.1	Numerical results from Task 5.12	67
5.2	Discounted expected value of the option	67
6.1	Measurement of a resistor	79
6.2	Measurement values for two groups	81
6.3	Estimated resistances from measurement groups with \hat{R}_{SA} in blue, \hat{R}_{EV} in red and \hat{R}_{LS} in green.	82
6.4	Observed probability density functions for groups. From left to right $N = 1000$, $N = 10000$ and $N = 100000$ with \hat{R}_{SA} in blue, \hat{R}_{EV} in red and \hat{R}_{LS} in green.	82
6.5	Observed standard deviation for groups. From left to right $N = 1000$, $N = 10000$ and $N = 100000$ with \hat{R}_{SA} in blue, \hat{R}_{EV} in red and \hat{R}_{LS} in green.	83
6.6	Comparison of histograms for the current $i(\cdot)$	84
7.1	Sample measurements and estimation for Example 7.10	90
7.2	Sample measurements and estimation for Task 7.11	92
7.3	Sample measurements and estimation for Example ??	97
8.1	Block diagram of the state space system (8.2)	104
8.2	IMU measurement data from gyros and accelerometers for sudden strikes	111
8.3	IMU Kalman filter fusion results in comparison to single sensor family results	112

List of Definitions and Theorems

Definition 1.1 Probability space	7
Definition 1.2 Random variable	7
Definition 1.3 Expected value or mean	8
Definition 1.4 Moment	8
Definition 1.5 Covariance	8
Definition 1.6 Probability density function	9
Definition 1.7 Gaussian (or normal) distribution	9
Definition 1.8 Mean square convergence	10
Definition 1.9 Ordinary Differential Equation	11
Definition 1.10 Initial Value Problem	11
Definition 1.11 Lipschitz Condition	12
Theorem 1.12 Existence and Uniqueness	12
Definition 1.13 Stochastic differential equation	13
Definition 1.14 Wiener process	13
Definition 2.1 Discrete time population dynamics	17
Definition 2.2 Continuous time population dynamics	18
Definition 2.3 Simple growth model	19
Definition 2.6 Logistic growth	21
Definition 2.7 Equilibrium	21
Corollary 2.8 Equilibrium	22
Theorem 2.9 Equilibrium	22
Corollary 2.10 Behavior	22
Definition 2.12 Exponential Stability	22
Theorem 2.14 Exponential Stability	23
Corollary 2.15 Convergence	24
Definition 2.17 Lotka-Volterra model	25
Corollary 2.18 Equilibria and convergence	26
Definition 2.20 Periodicity	27
Definition 2.21 Lotka-Volterra model with limited resources	28
Corollary 2.22 Equilibria and convergence	28

Definition 2.23 Generalized Lotka-Volterra model	29
Theorem 3.1 Newton's 3rd law	37
Definition 3.3 Constraints	40
Definition 3.5 Generalized coordinates	40
Definition 3.7 Generalized velocities	41
Definition 3.9 Generalized kinetic energy	42
Definition 3.11 Generalized forces	42
Definition 3.12 Generalized potential energy	43
Definition 3.14 Lagrangian	43
Theorem 3.15 Lagrangian equation	44
Definition 4.1 Network/graph	47
Definition 4.3 Undirected bond graph	48
Definition 4.6 Directed bond graph	50
Definition 4.7 Junction structure	51
Definition 4.8 0-junction	51
Definition 4.9 1-junction	52
Definition 4.10 Signal port	52
Definition 4.11 Internal bond	53
Definition 4.12 Simple junction structure	53
Definition 4.13 External bond	53
Definition 4.14 Two-port transformer	53
Definition 4.16 Two-port gyrator	54
Definition 4.18 General junction structure	55
Definition 4.19 Weighted junction structure	56
Definition 4.20 Environmental elements	56
Definition 4.21 1-port C energy storage	56
Definition 4.22 1-port I energy storage	56
Definition 4.23 1-port resistor	57
Definition 5.1 Terms of options	62
Corollary 5.2 Value of an option	62
Definition 5.4 Risk free payoff	63
Theorem 5.5 Value of option	63
Corollary 5.6 Value of option	64
Definition 5.7 Geometric Brownian motion	64
Theorem 5.8 Uniqueness Brownian motion	64
Corollary 5.10 Expected value and variance	65
Definition 6.3 Unbiased estimator	76

Definition 6.4 Weak and strong consistency	76
Definition 6.5 Efficiency	77
Theorem 6.6 Cramer-Rao rule	77
Corollary 6.8 Efficiency	78
Definition 7.1 Input-output model	85
Definition 7.3 Error variable	86
Definition 7.4 Least Square estimator	86
Definition 7.6 Linear-in-parameter input-output system	87
Definition 7.7 Linear least square estimation problem	88
Theorem 7.8 Solution of linear least square estimator	88
Corollary 7.12 Unbiasedness of the linear least square estimator	93
Corollary 7.13 Covariance of the linear least square estimator	94
Definition 7.17 Weighted Least Square estimator	96
Theorem 7.18 Solution of the weighted linear least square estimator	96
Corollary 7.20 Unbiasedness of the weighted linear least square estimator	98
Corollary 7.21 Covariance of the weighted linear least square estimator	98
Corollary 7.22 Minimal covariance of the weighted linear least square estimator	99
Corollary 8.1 Recursive estimation of mean	102
Definition 8.2 Filtering	104
Theorem 8.5 Mean propagation	106
Theorem 8.6 Covariance propagation	106
Theorem 8.9 Kalman filter for LTI systems without external input	109
Theorem 8.12 Kalman filter for LTI systems with external input	110

CHAPTER 1

INTRODUCTION

In this chapter we give a brief introduction to modelling and identification. We do this by using a simple example to model and illustrate the pitfalls associated with a model built from noisy measurements. In addition, we give a recap of terms from differential equations and probability theory that we will need throughout the lecture.

1.1. What is a “model”?

The first part of the lecture is about modelling. Intuitively, we all know what a model is. We have identified it by learning to control our actions using predictions about the effects of those actions. These predictions are based on a model and form a model of reality in our mind. There are simple connections, e.g. “I push a ball, then it rolls”. We can also build up very complicated systems, such as cars, supply chains or weather forecasts. The model is therefore something deterministic, without uncertainty and predictable for all time.

Unfortunately, as we have all experienced, models do not represent reality one to one. So when we use a model, there can be deviations between the model’s prediction and reality, especially over long time horizons. The reason for this is as follows: In a model, we always focus on the aspects we are interested in and do not try to describe all of reality. The problem is therefore divided into two parts,

- the model, which describes what we are interested in, and
- the environment, which contains everything else.

Since we cannot tell anything about the environment (because it is not modeled), interactions between model and environment can only be interpreted as disturbances. As we will see in the lecture, disturbances can also be modeled/considered and thus estimated.

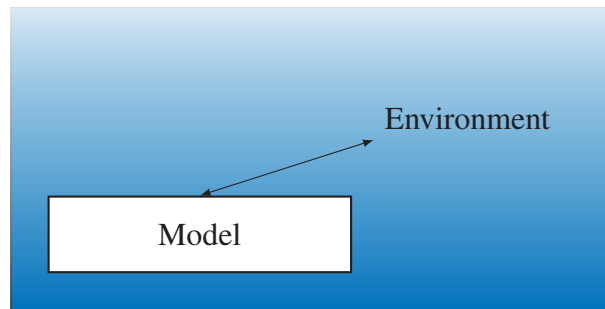


Figure 1.1.: Model and environment

During the modeling process, six principles need to be met:

1. Principle of Correctness: A model needs to present the facts correctly regarding structure and dynamics (semantics). Specific notation rules have to be considered (syntax).
2. Principle of Relevance: All relevant items have to be modeled. Non-relevant items have to be left out, i.e. the value of the model doesn't decline if these items are removed.
3. Principle of Cost vs. Benefit: The amount of effort to gather the data and produce the model must be balanced against the expected benefit.
4. Principle of Clarity: The model must be understandable and usable. The required knowledge for understanding the model should be as low as possible.
5. Principle of Comparability: A common approach to modeling ensures future comparability of different models that have been created independently from each other.
6. Principle of Systematic Structure: Models produced in different views should be capable of integration. Interfaces need to be designed to ensure interoperability.

This raises an interesting point: Since the modeler and the model user are typically different entities with different perspectives on the process, a good model for the modeler may be very different from a good model for the model user. For example, a detailed model may reflect reality very well, but it may be too complex to evaluate in real time and therefore not usable for feedback control. Therefore, modelling must be fit for use and the quality of a model is determined by the degree to which it meets the needs of the model user ("fit for use").

Combined, the aim of modeling theory is the following:

Modeling theory provides a systematic approach to mathematically describe those part of the problem, which are sufficient for its fitness of usage, and acknowledge the necessary conditions for its development.

In this lecture we focus on the quantitative description of a model, i.e. qualitative results such as “a ball will roll downhill” are not the kind of model properties we are looking for. Instead, we use laws, e.g. from physics or econometrics, to describe at least part of our impression of reality. We will also introduce and discuss properties of systems such as stability and observability. In general, the (mathematical) description of models varies depending on the time, space and amplitude properties considered. Figure 1.2 gives a rough overview of these properties.

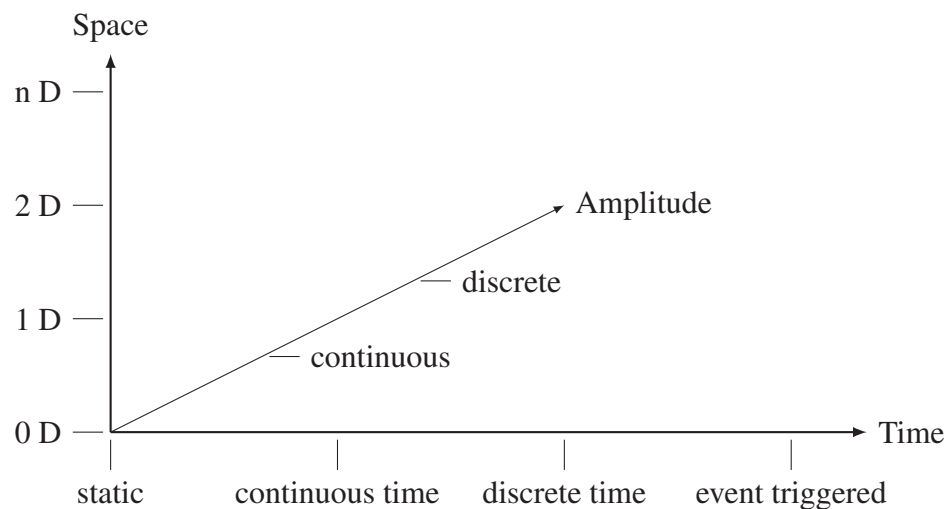


Figure 1.2.: Dimensions of model characteristics

These model are subject to parameters, which we aim to estimate later using respective data.

- Regarding time, we start off with static models, which are characterized by the fact that inputs, outputs and measurements of the system are available. In contrast to that, continuous time models exhibit data streams being received continuously. Discrete time models differ from that by the availability of data, which is received at certain, not necessarily equidistant time instances. Last, event triggered models require issues to trigger receiving data.
- Regarding space, models may vary from a simple connection to complex systems.
- Regarding amplitude, models may differ regarding continuous spaces like e.g. mass and discrete spaces such as gear shifts.

In general, a dynamic system can be seen as a blackbox, which assigns an output sequence y to a given input sequence u , cf. Figure 1.3.

Dynamic systems are categorized using a variety of properties, which allow for improved treatment of dynamic systems. Here, we only focus on the main research lines displayed in Table 1.1 below.

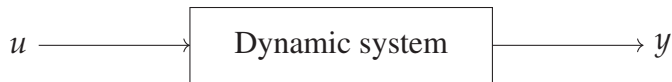


Figure 1.3.: General structure of a dynamic system

Simple type	Complex type
Linear The system is linear in the input and output variable.	Nonlinear The system may not be linear in either the input or the output variable.
Time invariant All parameters are constants	Time varying At least one parameter is time dependent.
Continuous time Time is given by a real valued variable.	Discrete time Time is given by sampling instants.
Input output model Input is directly mapped to output.	State space model Input triggers changes of an internal state, output is a linear combination of these internal states.
Deterministic None of the variables/parameters is a random variable.	Stochastic At least one variable or parameter is a random variable.

Table 1.1.: Division lines for dynamic systems

Within the lecture, we consider models satisfying the so called nonlinear discrete time control systems form

$$\begin{aligned} x(k+1) &= f(x(k), u(k)), \\ y(k) &= h(x(k)) \end{aligned} \tag{1.1}$$

or the continuous time form

$$\begin{aligned} \dot{x}(t) &= f(x(t), u(t)), \\ y(t) &= h(x(t)) \end{aligned} \tag{1.2}$$

where x represents the internal state of the system, u the external force on the system, f the law

or dynamics of the system, y the measured output, h the output or measurement function and k, t the discrete and continuous time respectively.

In particular, we will focus on modeling aspects of

- Growth Processes,
- Mechanical Processes,
- Electrical Processes, and
- Financial Processes.

Each of these topics is so large that we cannot cover them all. We therefore limit ourselves to certain aspects of these topics. With regard to growth processes, we will consider a biological model that can be used to describe the growth of a market for a product or the population of a species or several competing species. For mechanical processes, we will use the laws of motion to develop modular models of mechanical processes and introduce the Lagrangian and Hamiltonian approaches. For electrical processes, we consider the modeling of circuits and discuss the Bond graph approach. Finally, for financial applications, we will focus on option pricing and consider the Black-Scholes approach.

1.2. What is “identification”?

In the second part of the lecture we will use a given model and discuss methods to show its validity. To do this, our task is to match the behaviour of the model to that of the real process, which is also called fitting. To do this, we need not only the model itself, but also data from the process and a way to simulate the model. The fit of the simulated data to the real data is then qualified by an optimisation criterion, which, as described above, is defined by the degree to which the model satisfies the requirements of the model user. This criterion allows us to fit the model „best“ in the sense of the criterion. Finally, the model should always be validated, i.e. tested for failure or inconclusive results. Thus, any identification process consists of a series of basic steps:

1. Collect information on the system
2. Select a model to represent the system
3. Choose an optimization criterion
4. Fit the model parameters to the measurements accordingly
5. Validate the computed model

Note that some of the steps may be hidden from the user, or selected without the user being aware of a choice, which may lead to suboptimal or even poor performance. Unfortunately, fitting laws or models to observations creates new problems:

- First, we consider noisy measurements. In this context, noisy means that when we make a measurement, e.g. of length, weight, time, etc., errors occur because the instruments we use are not perfect.
- secondly, our laws and models are imperfect because reality is much more complex than the rules we apply. They also exhibit stochastic behavior, which makes it impossible to predict their outcome accurately.

To identify the system, we split the model into a deterministic and a stochastic part. The deterministic aspects are captured by the mathematical system model. These are complemented by the stochastic behavior, which is modeled as a noise distortion. The aim of identification theory is therefore as follows:

Identification theory provides a systematic approach to fit the mathematical model to the deterministic part as well as possible, and to eliminate the noise distortions as much as possible.

Within this lecture, we particularly focus on the techniques of the

- Least Square Estimator, and of the
- Kalman Filter.

Note that the terms estimator and filter are similar, yet an estimator refers to a static problem and a filter to a dynamic one. Still, estimators can be applied to dynamical problem, but are not ideally suited.

1.3. Stochastic parameters

Within identification, we use a finite number of possibly noisy measurements to compute parameters within the model. These are therefore stochastic variables. To fully characterise such a variable, we need its probability density function. In practice, it is very difficult to derive this function, but it can be described by a few numbers, i.e. the mean and the covariance, which can be thought of as the location and dispersion of the estimate.

To introduce these numbers formally, we first need the notion of a probability space:

Definition 1.1 (Probability space).

Consider a set Ω , a set of subsets $\mathcal{F} \subseteq 2^\Omega$ and a function $P : \mathcal{F} \rightarrow [0, 1]$. Then, we call the triple (Ω, \mathcal{F}, P) a *probability space* if

- the sample space Ω is a non-empty set,
- the σ -algebra \mathcal{F} of events satisfies
 - \mathcal{F} contains the empty set, i.e.

$$\emptyset \in \mathcal{F},$$

- \mathcal{F} is closed under complements, i.e.

$$A \in \mathcal{F} \implies \Omega \setminus A \in \mathcal{F},$$

- \mathcal{F} is closed under countable unions, i.e.

$$A_i \in \mathcal{F} \forall i \in \{1, 2, \dots, k\}, k < \infty \implies \bigcup_{i \in \{1, 2, \dots, k\}} A_i \in \mathcal{F}$$

- the probability measure P satisfies
 - P is countably additive, i.e.

$$A_i \in \mathcal{F} \forall i \in \{1, 2, \dots, k\}, k < \infty \text{ with } A_i \cap A_j = \emptyset \forall i, j \in \{1, 2, \dots, k\}, i \neq j$$

$$\implies P \left(\bigcup_{i \in \{1, 2, \dots, k\}} A_i \right) = \sum_{i \in \{1, 2, \dots, k\}} P(A_i),$$

- the measure of the sample space Ω is one, i.e.

$$P(\Omega) = 1.$$

In short, a probability space is a measure space, but with the additional property that the measure of the whole space is equal to one. Secondly, we require so called random variables:

Definition 1.2 (Random variable).

Consider a probability space (Ω, \mathcal{F}, P) and a measurable space E with σ -algebra \mathcal{E} of E . Then

we call a function $X : \Omega \rightarrow E$ a *random variable* if

$$\forall B \in \mathcal{E} : X^{-1}(B) \in \mathcal{F}, \quad \text{where } X^{-1}(B) := \{\omega \in \Omega \mid X(\omega) \in B\}.$$

Hence, a random variable is a function, which allows us to use a more comfortable description of properties or measurements of a sample, i.e. if B is an interval $[a, b]$ or the property “lottery player”, then we identify the corresponding event $X^{-1}(B)$ in the σ -algebra \mathcal{F} .

Now, we can introduce the expected value, sometimes also called mean, first moment or expectation:

Definition 1.3 (Expected value or mean).

Consider a probability space (Ω, \mathcal{F}, P) and a random variable X defined on that triple. Then, the *expected value* $E(X)$ or mean of X is defined as the Lebesgue integral

$$E(X) := \int_{\Omega} X dP = \int_{\Omega} X(\omega) dP(\omega) \quad (1.3)$$

whenever the integral exists.

Note that since the integral may not converge absolutely, not all random variables have a finite expected value, and for some it is not defined at all (e.g., Cauchy distribution).

In order to define the second important number, the covariance, we first introduce the notion of moments:

Definition 1.4 (Moment).

Consider a probability space (Ω, \mathcal{F}, P) , a natural number $n \in \mathbb{N}$ and a random variable X defined on that triple. Then, the n -th *moment* is given by

$$m_n := E(X^n). \quad (1.4)$$

Hence, the mean is also the first moment. Regarding the covariance, we require the second moment to describe, how much two random variables in one probability space change together, i.e. what the nature of their connection and how strong this connection is:

Definition 1.5 (Covariance).

Consider a probability space (Ω, \mathcal{F}, P) and two random variables X and Y defined on that triple.

Then, the *covariance* $\text{Cov}(X, Y)$ is defined as

$$\text{Cov}(X, Y) := E((X - E(X))(Y - E(Y))) \quad (1.5)$$

whenever the second moments of X and Y exist.

If $X = Y$, then covariance is called *variance* and we obtain $\text{Cov}(X, X) = \sigma^2(X)$.

Higher moments describe the skewness and curtosis of the probability function P , which can be interpreted as a measure of deviation from a normal distribution and a measure of deviation from a symmetric distribution, respectively.

The following notion of a so-called probability density function uses the nice property of a random variable to be a transformation into an easily interpretable space. I.e. it describes the relative likelihood for this random variable to take a given value (evaluated in the image space of the random variable):

Definition 1.6 (Probability density function).

Consider a probability space (Ω, \mathcal{F}, P) and a random variable $X : \Omega \rightarrow E$ defined on that triple, where the set E equipped with measure μ and \mathcal{E} is a σ -algebra of E . Then, any measurable function $f : \mathcal{E} \rightarrow \mathbb{R}_0^+$, which satisfies

$$\Pr(X \in B) \left(= \int_{X^{-1}(B)} dP \right) = \int_B f d\mu \quad (1.6)$$

for any measurable set $B \in \mathcal{E}$ is called a *probability density function*.

One of the most famous probability density functions induces the so called *Gaussian* random variables.

Definition 1.7 (Gaussian (or normal) distribution).

Consider a probability space (Ω, \mathcal{F}, P) and a random variable $X : \Omega \rightarrow E$ defined on that triple, where the set E equipped with measure μ and \mathcal{E} is a σ -algebra of E . Suppose that the parameters $\mu, \sigma \in \mathbb{R}$ with $\sigma > 0$ define the density function

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp^{-\frac{(x-\mu)^2}{2\sigma^2}}. \quad (1.7)$$

of the random variable X . Then X is called a *Gaussian random variable*, also written $X \in \mathcal{N}(\mu, \sigma^2)$, and f is called *Gaussian distribution*.

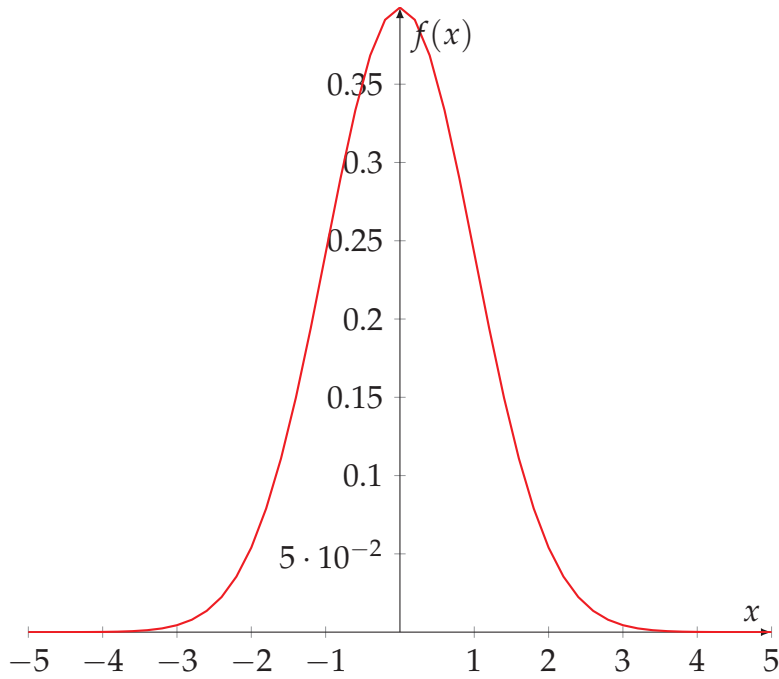


Figure 1.4.: Sketch of the Gaussian distribution

Last, we require that the identification methods somehow produce a solution, which converges towards the true parameter of the system. Here, we solely focus on the concept of mean square convergence.

Definition 1.8 (Mean square convergence).

Consider a probability space (Ω, \mathcal{F}, P) and a sequence of random variables $X(N)$, $N \in \mathbb{N}$ and a random variable X , both defined on that triple. Then, we call $X(N)$ **to converge to X in mean square** if

- $E(|X|^2) < \infty$,
- $E(|X(N)|^2) < \infty$ for all $N \in \mathbb{N}$, and
- $\lim_{N \rightarrow \infty} E(|X(N) - X|^2) = 0$.

For short, we write $\text{l.i.m.}_{N \rightarrow \infty} X(N) = X$.

As this concept is based on distinct properties of the random variables, i.e. of its first and second moment, it is readily checkable within the identification process.

1.4. Models

Starting with deterministic and continuous time case, we consider ordinary differential equations:

Definition 1.9 (Ordinary Differential Equation).

An *ordinary differential equation* in \mathbb{R}^{n_x} , $n_x \in \mathbb{N}$, is given by

$$\frac{d}{dt}x(t) = f(t, x(t)) \quad (1.8)$$

where $f : \mathbb{D} \rightarrow \mathbb{R}^{n_x}$ is a continuous function and \mathbb{D} is an open subset of $\mathbb{R} \times \mathbb{R}^{n_x}$.

The solution of (1.8) is a continuously differentiable function $x : \mathbb{R} \rightarrow \mathbb{R}^{n_x}$, which satisfies (1.8). In general, we will use the following denomination throughout the script:

- The independent variable t is referred to as time, although other interpretations are possible.
- Instead of $\frac{d}{dt}x(t)$ we will often use the abbreviation $\dot{x}(t)$.
- The function $x(t)$ is called solution or trajectory.
- If the function f is independent of t , i.e. $\dot{x}(t) = f(x(t))$, then the differential equation is called autonomous.

An ordinary differential equation typically possesses infinitely many solutions. To obtain a unique solution, we have to introduce a constraint, the so called initial value constraint. Combined with the differential equation (1.8), this reveals the so called initial value problem:

Definition 1.10 (Initial Value Problem).

Consider values t_0 and $x_0 \in \mathbb{R}^{n_x}$ to be given. Then the *initial value problem* is to find the solution satisfying the differential equation

$$\dot{x}(t) = f(t, x(t)) \quad (1.8)$$

and the *initial value condition*

$$x(t_0) = x_0. \quad (1.9)$$

Here, the time $t_0 \in \mathbb{R}$ is called initial time and the value $x_0 \in \mathbb{R}^{n_x}$ is called initial value. Both the pair (t_0, x_0) and equation (1.9) are called initial condition.

Under certain conditions, existence and uniqueness of a solution to the problem from Definition 1.10 can be shown. This is the so called Lipschitz condition

Definition 1.11 (Lipschitz Condition).

Consider a function $f : \mathbb{D} \rightarrow \mathbb{R}^{n_x}$ with $\mathbb{D} \subset \mathbb{R} \times \mathbb{R}^{n_x}$. Then f is called *Lipschitz* in its second argument, if for each compact set $K \subset \mathbb{D}$ there exists a constant $L > 0$ and

$$\|f(t, x) - f(t, y)\| \leq L\|x - y\| \quad (1.10)$$

holds for all $t \in \mathbb{R}$ and all $x, y \in \mathbb{R}^{n_x}$ with $(t, x), (t, y) \in K$.

Using this property, we can show the following:

Theorem 1.12 (Existence and Uniqueness).

Consider a differential equation (1.8) with $f : \mathbb{D} \rightarrow \mathbb{R}^{n_x}$ and $\mathbb{D} \subset \mathbb{R} \times \mathbb{R}^{n_x}$. Moreover, f is considered to be continuous and Lipschitz continuous in the second argument. Then for each initial condition $(t_0, x_0) \in \mathbb{D}$, there exists a unique solution $x(t; t_0, x_0)$ of the initial value problem (1.8), (1.9). This solution is defined for all t from an open maximal interval of existence I_{t_0, x_0} with $t_0 \in I_{t_0, x_0}$.

Here, we like to note that the dynamic reveals a *flow* of the system at hand, whereas a *trajectory* is bound to a specific initial value and input sequence. The following Figure 1.5 illustrates the idea of flow and trajectory. In this case, the flow is colored to mark its intensity whereas the arrows point into its direction. The trajectory is evaluated for a specific initial value and „follows“ the flow accordingly.

Note that at the boundary of the interval of existence I_{t_0, x_0} the solution ceases to exist. If the interval is bounded, then there are two possible reasons for that: For one, the solution may diverge, or secondly the solution converges to a boundary point of \mathbb{D} . In the remainder of this script, we will always assume that the assumptions of Theorem 1.12 are met without explicitly stating it.

Table 1.2.: Advantages and disadvantages of differential equations

Advantage	Disadvantage
✓ Allows to model dynamics	✗ May require solvers
✓ Allows validation of properties analytically	✗ Requires identification

To capture randomness within a model, we can extend ordinary differential equations to so called *stochastic differential equations*.

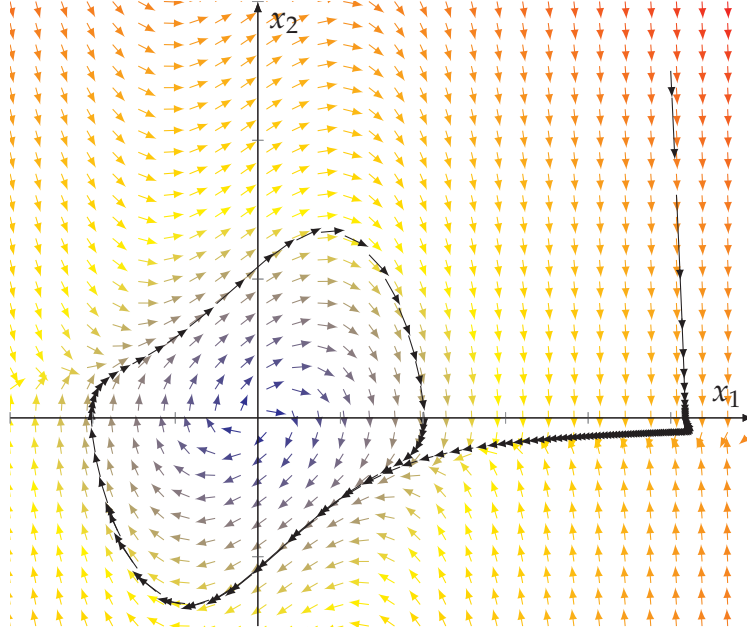


Figure 1.5.: Sketch of a dynamic flow and a trajectory

Definition 1.13 (Stochastic differential equation).

Consider deterministic functions $a, b : \mathbb{R} \times \mathbb{R}^{n_x} \rightarrow \mathbb{R}^{n_x}$, a probability space (Ω, \mathcal{F}, P) and a random variable $X : \mathbb{R} \times \Omega \rightarrow \mathbb{R}^{n_x}$ to be given. Then we call

$$\dot{x}(t) = a(t, x(t)) + b(t, x(t))X(t, \cdot) \quad (1.11)$$

a *stochastic differential equation*.

Here, the introduction of the random variable X causes possibly multiple solutions to exist. Since the realization of the random variable $X(\cdot, \omega)$ depends on chance, the solution also depends on chance. In turn, once the realization $\omega \in \Omega$ is fixed, (1.11) is an ordinary differential equation with a unique solution, i.e. for each realization which is also called a *path*, there exists one solution.

Here, we have a more close look at a specific path, the so called *Wiener process*.

Definition 1.14 (Wiener process).

Consider a probability space (Ω, \mathcal{F}, P) and a random variable $W : \mathbb{R} \times \Omega \rightarrow \mathbb{R}^{n_x}$ to be given. We call W a *Wiener process* if the following conditions are satisfied:

1. $W(t, \cdot)$ is a Gaussian random variable with $E(W(t, \cdot)) = 0$ and $\sigma^2(W(t, \cdot)) = t$.

2. For $t_1 \geq t_0 \geq 0$ the *increments* $W(t_1, \cdot) - W(t_0, \cdot)$ are Gaussian random variables with $E(W(t_1, \cdot) - W(t_0, \cdot)) = 0$ and $\sigma^2(W(t_1, \cdot) - W(t_0, \cdot)) = t_1 - t_0$.
3. For $t_3 \geq t_2 \geq t_1 \geq t_0 \geq 0$ the increments $W(t_3, \cdot) - W(t_2, \cdot)$ and $W(t_1, \cdot) - W(t_0, \cdot)$ are Gaussian random variables.

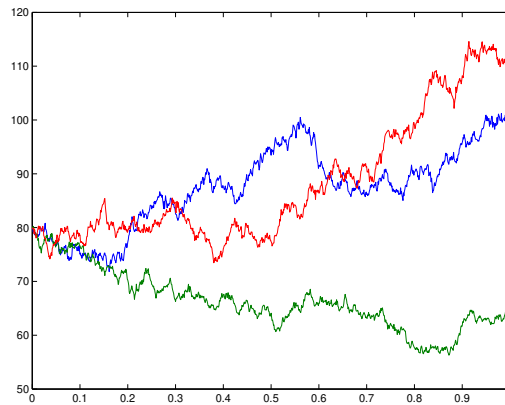


Figure 1.6.: Different paths of a Wiener process

A path $W(t, \omega)$ of W is one of many possible arbitrary functions, which (in the whole) satisfy the conditions above. Indeed, one can show that these paths are almost surely continuous in t , i.e. the event $A = \{\omega \in \Omega \mid X(t, \omega) \text{ is continuous in } t\}$ exhibits probability $\Pr(A) = 1$, and almost surely nowhere differentiable.

We like to point out that by condition 3, a Wiener process is memory free, and paths could at any time move upwards and downwards with exactly the same probability, no matter the past development.

Table 1.3.: Advantages and disadvantages of stochastic processes

Advantage	Disadvantage
✓ Allows to model uncertainty	✗ Converges in stochastic sense
✓ Extends ordinary differential equations	✗ Requires extensive simulation

The Wiener process will form the basis of the so called Monte-Carlo method, which we will introduce and discuss in the upcoming Chapter 5. Yet first, we will start with growth, mechanical and electric processes in Chapters 2, 3 and 4 respectively.

Part I.

Modeling

CHAPTER 2

GROWTH PROCESSES

Deterministic models cover a range of applications such as growth, production and transport processes, as well as biological/chemical reactions or the spread of disease or information. These models can also be used in different areas such as market forecasting or product displacement. One of the classic systems for modelling growth processes is also called the logistic equation. In this chapter we will stick to the classical applications and analyse in more detail models with one or more entities, with and without resource constraints.

2.1. Growth dynamics for one object type

The analysis of growth is called population dynamics in the biological case. Within this section, we concentrate on one object type and analyze this problem in detail.

2.1.1. From difference to differential equation

Dynamics of objects are discrete in nature: The size of a set of objects is usually measured by the number of individual objects, which is a natural number. Similarly, measurements are typically taken at discrete instances in time $t_1 < t_2 < \dots$. Consequently, this gives us a system of form

Definition 2.1 (Discrete time population dynamics).

Given

$\Delta B(t_k)$	Number of created objects in the time interval $[t_k, t_{k+1}]$
$\Delta D(t_k)$	Number of discreated objects in the time interval $[t_k, t_{k+1}]$
$\Delta M(t_k)$	Number of migrated objects in the time interval $[t_k, t_{k+1}]$

we call the system

$$x(t_{k+1}) = x(t_k) + \Delta B(t_k) - \Delta D(t_k) + \Delta M(t_k) \quad (2.1)$$

population dynamics.

Equations of type (2.1) are called difference equations. To be able to apply analytical tools for ordinary differential equations, we have to modify both the state axes and the time axes to be (in this case non negative) reals.

To obtain a differential equation from (2.1), we assume that all time instances t_k are equally distributed, i.e. $t_{k+1} - t_k =: \Delta t$ for all $k \in \mathbb{N}$. Hence, we obtain

$$\frac{x(t + \Delta t) - x(t)}{\Delta t} = \frac{\Delta B(t)}{\Delta t} - \frac{\Delta D(t)}{\Delta t} + \frac{\Delta M(t)}{\Delta t}.$$

Note that ΔB , ΔD and ΔM depend on Δt , even if this is not explicitly mentioned in our notation.

Letting $\Delta t \rightarrow 0$, we obtain

Definition 2.2 (Continuous time population dynamics).

Suppose a population dynamics (2.1) to be given. Then we call

$$\dot{x}(t) = b(t) - d(t) + m(t). \quad (2.2)$$

with

$$b(t) = \lim_{\Delta t \rightarrow 0} \frac{\Delta B(t)}{\Delta t}, \quad d(t) = \lim_{\Delta t \rightarrow 0} \frac{\Delta D(t)}{\Delta t} \quad \text{and} \quad m(t) = \lim_{\Delta t \rightarrow 0} \frac{\Delta M(t)}{\Delta t}.$$

continuous time population dynamics.

Proceeding this way would be a good idea if ΔB , ΔD and ΔM were known. Here, we do not follow this route but instead deduce b and d from model assumptions directly.

2.1.2. Simple growth model

The most simple growth model is given by the following assumptions:

1. The rate of creation is linearly proportional to the current size of the set of objects:

$$b(t) = \gamma x(t) \quad \text{for some } \gamma \in \mathbb{R}$$

2. The rate of decreation is linearly proportional to the current size of the set of objects:

$$d(t) = \sigma x(t) \quad \text{for some } \sigma \in \mathbb{R}$$

3. There is no migration:

$$m(t) \equiv 0$$

This leads to the following system

Definition 2.3 (Simple growth model).

Given a continuous time population dynamics (2.2) with $m(t) \equiv 0$, the differential equation

$$\dot{x}(t) = \lambda x(t) \tag{2.3}$$

is called *simple growth model* where $\lambda = \gamma - \sigma$ represents the growth rate as difference between birth and death rate. The solutions of (2.3) with initial condition $x(t_0) = x_0$ are given by

$$x(t; t_0 x_0) = x_0 \exp^{\lambda(t-t_0)}.$$

Note that $x(t)$ denotes the size of the set of objects. Hence, we can only allow for $x(t) \geq 0$, and in particular $x_0 \geq 0$. Here and in the following, we use the abbreviation $\mathbb{R}^+ = \{x \in \mathbb{R} \mid x > 0\}$ and $\mathbb{R}_0^+ = \mathbb{R}^+ \cup \{0\}$.

Although this model is very simple, it still describes some growth phenomena pretty well.

Task 2.4 (Population growth worldwide)

Figure 2.1 shows the size of the world population between 1950 and 2010 in billions. Fit a simple growth model to match the data.

Solution to Task 2.4: A respective solution of (2.3) can be obtained by values $x_0 = 2.5747$ and $\lambda = 0.0172$.

Task 2.5 (Population Europe)

Capture the development of the population development in Europe as shown in Figure 2.2 using a simple growth model.

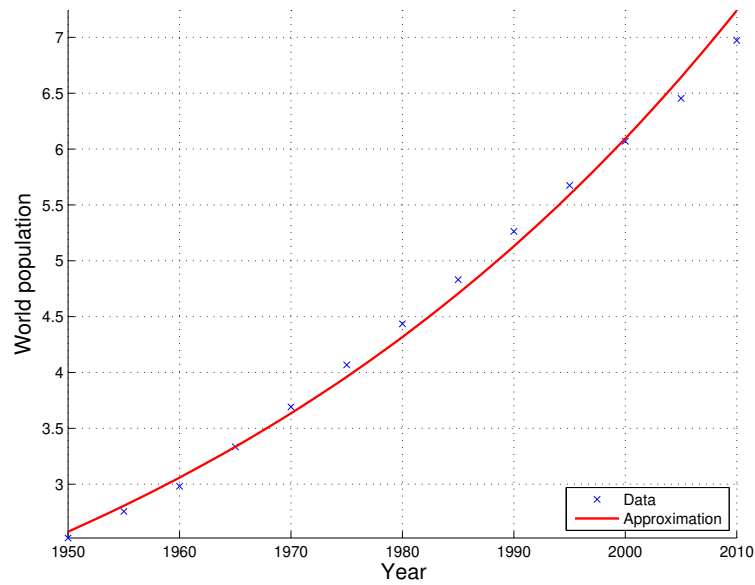


Figure 2.1.: Growth of the world population and solution of (2.3) for identified parameters

Solution to Task 2.5: The development of the stalling population in Europe cannot be captured correctly, cf. Figure 2.2.

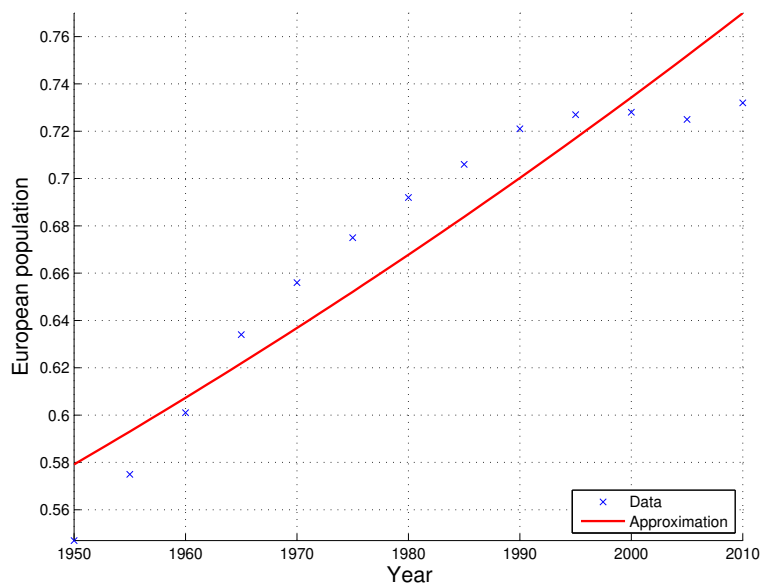


Figure 2.2.: Growth of the European population and solution of (2.3) for identified parameters

This inability arises since for $\lambda > 0$ we have that $\exp^{\lambda t} \rightarrow \infty$ as $t \rightarrow \infty$ which illustrates the necessity for capacity bounds.

2.1.3. Logistic growth model

To model such a slowed down growth, we integrate an upper bound $C > 0$ for the size of the set of objects in (2.3). C represents a capacity, which in Biology is subject to the available resources such as food, water etc., in production limiting factors may be machinery, workforce, space, and in logistics factors such as road capacities and time may reveal restrictions. In particular, we want to enforce

1. If $x < C$, then we have $g(x) > 0$ reflecting availability for growth.
2. If $x > C$, then we have $g(x) < 0$ reflecting negative growth.

The simplest function, which exhibits such a behavior, is the linear function $g(x) = C - x$. Applying this function, we obtain

Definition 2.6 (Logistic growth).

Consider a continuous time population dynamics (2.2) with capacity $C > 0$, then

$$\dot{x}(t) = \lambda (C - x(t)) x(t), \quad (2.4)$$

is also called *logistic growth* or *logistics equation*. The explicit solutions of (2.4) with initial condition $x(t_0) = x_0$ are given by

$$x(t; t_0, x_0) = \frac{C}{1 + \left(\frac{C}{x_0} - 1\right) \exp^{-\lambda C(t-t_0)}}. \quad (2.5)$$

Note that the expression $\lambda (C - x)$ is a nonlinear growth rate.

Based on the solution formula, we could analyze the behavior of the latter. To generalize our analysis, we first introduce some important terms for differential equations.

Definition 2.7 (Equilibrium).

A point $x^* \in \mathbb{R}^{n_x}$ is called equilibrium (or fixed point) of a differential equation (1.2) if $x(t; t_0, x^*) = x^*$ for all $t, t_0 \in \mathbb{R}$.

One can easily see that a point x^* is an equilibrium if and only if $f(t, x^*) = 0$ for all $t \in \mathbb{R}$.

Corollary 2.8 (Equilibrium).

Given the logistics growth model (2.4), x^* is an equilibrium if and only if $x^* = 0$ and $x^+ = C$.

Equilibria are of particular interest due to their potential in analyzing the long term behavior of solutions.

Theorem 2.9 (Equilibrium).

Consider differential equation (1.2) where f is autonomous. Moreover, the solution $x(t; t_0, x_0)$ converges to a point $x^* \in \mathbb{R}^{n_x}$ for $t \rightarrow \infty$ or $t \rightarrow -\infty$. Then x^* is an equilibrium.

Regarding model (2.4), we can see that solutions $x(t; t_0, x_0)$ are growing strictly monotone between the two equilibria, i.e. $\dot{x}(t) > 0$ if $x(t) \in (0, C)$, and $\dot{x}(t) < 0$ if $x(t) > C$. Since the solutions in positive time are bounded by the equilibrium solution $x(t) = x^+ = C$ and cannot intersect due to uniqueness, cf. Theorem 1.12, they are monotone and bounded, and therefore they converge.

Corollary 2.10 (Behavior).

Given the logistics growth model (2.4), we have that all solutions with

- $x(t_0) > 0$ converge to $x^+ = C$ for $t \rightarrow \infty$,
- $x(t_0) \in [0, C)$ converge to 0 for $t \rightarrow -\infty$, and
- $x(t_0) > C$ diverge to $x(t) \rightarrow \infty$ for $t \rightarrow -\infty$.

Remark 2.11

As a consequence of Theorem 2.9 we know that equilibria represent all possible limits of solutions in the autonomous case.

For higher dimensions, monotonicity can be substituted by the following:

Definition 2.12 (Exponential Stability).

Consider a differential equation (1.2).

1. An equilibrium $x^* \in \mathbb{R}^{n_x}$ is called (locally) exponentially stable, if there exists a neighborhood \mathcal{N} of x^* and parameters $\lambda, \theta > 0$ such that

$$\|x(t; t_0, x_0) - x^*\| \leq \theta \exp^{-\lambda(t-t_0)} \|x_0 - x^*\|$$

holds for all $x_0 \in \mathcal{N}$, $t_0 \in \mathbb{R}$ and all $t \geq t_0$.

2. An equilibrium $x^* \in \mathbb{R}^{n_x}$ is called exponentially unstable, if parameter $\lambda, \theta > 0$ and a neighborhood \mathcal{N} of x^* exist such that within each neighborhood $\mathcal{N}_0 \subset \mathcal{N}$ of x^* there exists a point $x_0 \in \mathcal{N}_0$ which satisfies

$$\|x(t; t_0, x_0) - x^*\| \geq \theta \exp^{-\lambda(t-t_0)} \|x_0 - x^*\|$$

for all $t \geq t_0$ for which $x(t; t_0, x_0) \in \mathcal{N}$ holds.

3. An equilibrium $x^* \in \mathbb{R}^{n_x}$ is called exponentially antistable, if parameter $\lambda, \theta > 0$ and a neighborhood \mathcal{N} of x^* exist such that for all $x_0 \in \mathcal{N}$ with $x_0 \neq x^*$ and all $t_0 \in \mathbb{R}$ the inequality

$$\|x(t; t_0, x_0) - x^*\| \geq \theta \exp^{-\lambda(t-t_0)} \|x_0 - x^*\|$$

for all $t \geq t_0$ for which $x(t; t_0, x_0) \in \mathcal{N}$ holds.

Hence, for $t \rightarrow \infty$ and Case 1, all solutions from a neighborhood \mathcal{N} of the equilibrium x^* converge to the equilibrium x^* . In Case 3, all solutions move away from x^* for growing t , i.e. convergence is not possible. In Case 2 there exist solutions which start arbitrarily close to x^* but move away from it. However, there may exist initial values $x_0 \neq x^*$, for which the solution $x(t; t_0, x_0)$ converges to x^* .

Remark 2.13

Note that Cases 1–3 do not describe all possible scenarios. For example, a function $\beta(\|x_0 - x^\|, t)$ may exist, which converges to zero slower than $\theta \exp^{-\lambda(t-t_0)} \|x_0 - x^*\|$ and that instead of Case 1 the inequality*

$$\|x(t; t_0, x_0) - x^*\| \leq \beta(\|x_0 - x^*\|, t)$$

holds.

The reason for choosing the definition of the (restricted case of) exponential estimates lies in the simplicity of checking these criteria — at least for the case of autonomous differential equations.

Theorem 2.14 (Exponential Stability).

Consider an equilibrium $x^ \in \mathbb{R}^{n_x}$ of a differential equation (1.2) with autonomous vector field*

$f : \mathbb{R}^{n_x} \rightarrow \mathbb{R}^{n_x}$. Suppose f is continuously differentiable in a neighborhood of x^* and that $Df(x^*) \in \mathbb{R}^{n_x \times n_x}$ represents the Jacobian of f at x^* . Then the following holds:

1. The equilibrium x^* is (locally) exponentially stable if and only if the real parts of all Eigenvalues $\lambda_i \in \mathbb{C}$ of $Df(x^*)$ are negative.
2. The equilibrium x^* is exponentially unstable if and only if there exists one Eigenvalue $\lambda_i \in \mathbb{C}$ of $Df(x^*)$ with positive real part.
3. The equilibrium x^* is exponentially antistable if and only if the real part of all Eigenvalues $\lambda_i \in \mathbb{C}$ of $Df(x^*)$ are positive.

Proofs for these results can be found in the book [2]. The Jacobian $Df(x^*)$ is also often called the *linearization* of (1.2) at x^* .

Regarding our logistics growth model, we see the following:

Corollary 2.15 (Convergence).

Given the logistics growth model (2.4), we have the equilibria $x^* = 0$ and $x^+ = C$ and

$$Df(x) = \lambda(C - x) - \lambda x \quad \Rightarrow \quad Df(x^*) = \lambda C > 0 \text{ and } Df(x^+) = -\lambda C < 0.$$

Hence, $x^* = 0$ is exponentially antistable and $x^+ = C$ is exponentially stable.

Task 2.16 (Solution of logistics growth model)

Sketch the solution of the logistics growth model (2.4) with $C = \lambda = 1$ for initial values $x_0 \in \{0, 0.1, 1, 2\}$.

Solution to Task 2.16: The solution is plotted in Figure 2.3.

Table 2.1.: Advantages and disadvantages of logistics growth model

Advantage	Disadvantage
✓ Allows complete analysis	✗ Limited to one species
✓ Allows in-/decrease and limitation	✗ Unfit for spatial distribution
Continued on next page	

Table 2.1 – continued from previous page

Advantage	Disadvantage
✓ Focuses on mass and flow	✗ Unable to specify unit behavior

2.2. Lotka-Volterra model

In this section we extend the model (2.3) to the case of several types of objects. To this end, we first focus on the case with two objects where the first one represents a supply source for the second one. The first one is often referred to as prey or supply while the second one is called predator or production/consumer. The case of limited resources (2.4) can be treated similarly.

To extend our model (2.3) to two types of objects, we denote the set of objects of the first by x_1 and of the second by x_2 . For our model, we make the following assumptions:

1. The set of objects x_1 grows according to (2.3) with $\lambda = \gamma - \sigma$. Here, the rate of creation γ is constant and the rate of decreation is given by $\sigma = \tilde{\sigma} + bx_2$. The rate of decreation consists of a constant term $\tilde{\sigma} \in (0, \gamma)$ representing the natural decreation, and a proportional term bx_2 representing the absorption by x_2 . Hence, for $x_2 = 0$ the set of objects x_1 grows exponentially. Here, we set $a = \gamma - \tilde{\sigma}$.
2. The set of objects x_2 also evolves according to (2.3) with $\lambda = \gamma - \sigma$. Here, the rate of decreation σ is constant and the rate of creation $\gamma = \tilde{\gamma} + dx_1$ consists of the natural rate of creation $\tilde{\gamma} \in (0, \sigma)$ and a proportional term with cofactor $d > 0$. Hence, the rate of creation is affine linearly depending on the number of objects x_1 . For $x_1 = 0$ the set of objects x_2 is dying out as $\sigma > \tilde{\gamma}$. Here, we set $c = \sigma - \tilde{\gamma}$.

Combined, we obtain the following

Definition 2.17 (Lotka-Volterra model).

The system

$$\begin{aligned}\dot{x}_1(t) &= ax_1(t) - bx_1(t)x_2(t) \\ \dot{x}_2(t) &= -cx_2(t) + dx_1(t)x_2(t)\end{aligned}\tag{2.6}$$

with parameters $a, b, c, d > 0$ is called *Lotka–Volterra model*.

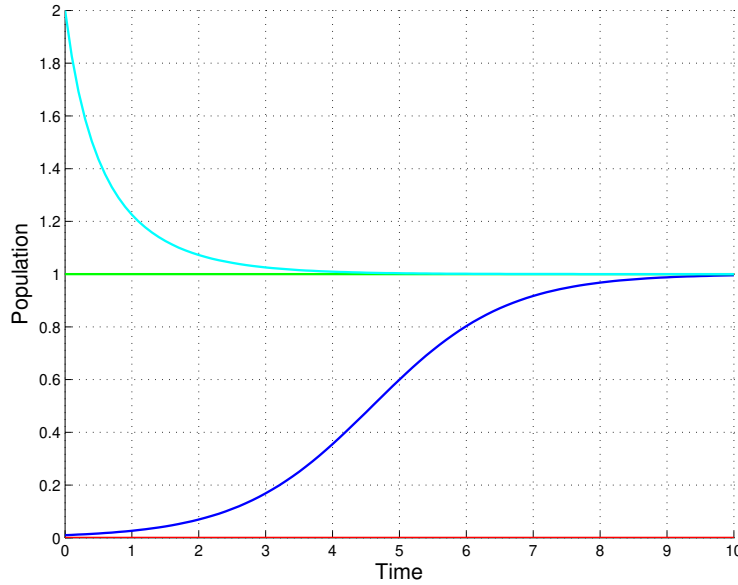


Figure 2.3.: Solutions of the Lotka-Volterra model (2.4)

For the analysis of (2.6), we first reduce the number of parameters. To this end, we apply the coordinate transformations $x_1 \rightarrow \frac{d}{c}x_1$ and $x_2 \rightarrow \frac{b}{a}x_2$, which gives us

$$\begin{aligned}\dot{x}_1(t) &= ax_1(t)(1 - x_2(t)) \\ \dot{x}_2(t) &= -cx_2(t)(1 - x_1(t))\end{aligned}\tag{2.7}$$

Utilizing our approach from the last section, we directly obtain

Corollary 2.18 (Equilibria and convergence).

The equilibria of the Lotka-Volterra model (2.7) are given by $x^* = (0, 0)^\top$ and $x^+ = (1, 1)^\top$ with Jacobian

$$Df(x^*) = \begin{pmatrix} a(1 - x_2^*) & -ax_1^* \\ cx_2^* & -c(1 - x_1^*) \end{pmatrix} = \begin{pmatrix} a & 0 \\ 0 & -c \end{pmatrix} \quad \text{and} \quad Df(x^+) = \begin{pmatrix} 0 & -a \\ c & 0 \end{pmatrix}.$$

For x^* the Eigenvalues are a and $-c$ rendering the point unstable. For x^+ , the Eigenvalues are $\pm\sqrt{-ca}$ rendering the point to be neither stable nor unstable.

Task 2.19 (Solution of Lotka-Volterra growth model)

Sketch the solution of the Lotka-Volterra model (2.7) with $a = c = 1$.

Solution to Task 2.19: The solution is plotted in Figure 2.4.

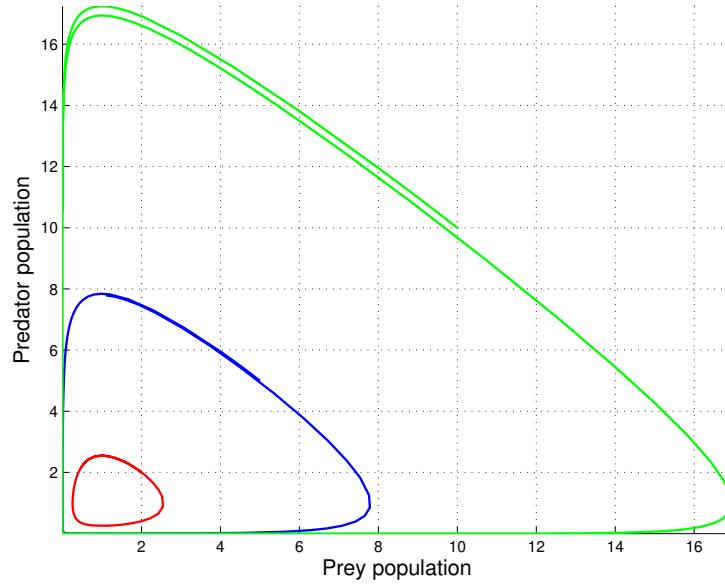


Figure 2.4.: Solutions for the Lotka-Volterra model (2.7) with $a = c = 1$

All solutions are moving along periodic orbits around x^+ . More formally, we can state the following:

Definition 2.20 (Periodicity).

A solution $x(t; t_0, x_0)$ is called periodic, if there exists a $T > 0$ such that

$$x(t; t_0, x_0) = x(t + T; t_0, x_0)$$

holds for all $t \in \mathbb{R}$. The time T is called the period of the solution.

In particular, the solution of an autonomous differential equation is periodic if and only if there exist two time instances $t_1 < t_2 \in \mathbb{R}$ such that $x(t_1) = x(t_2) = x_p$.

Similar to the logistics growth model, we can extend the Lotka-Volterra model to include limited resources. To this end, we modify our model assumption 1 as follows:

- 1'. The set of objects x_1 evolves according to (2.3) with $\lambda = \gamma - \sigma$ and there exists a bounding rate $e > 0$. Here, the rate of creation γ and the bounding rate e are constant and the rate of decreation is given by $\sigma = \tilde{\sigma} + bx_2$. There only exist bounded resources for x_1 and the

rate of decreation consists of a constant term $\tilde{\sigma} \in (0, \gamma)$ representing the natural rate of decreation, and a proportional term representing the absorption by bx_2 . Hence, for $x_2 = 0$ the set of objects x_1 approaches $C = a/e$ with $a = \gamma - \tilde{\sigma}$.

Hence, we obtain

Definition 2.21 (Lotka-Volterra model with limited resources).

We call the system

$$\begin{aligned}\dot{x}_1(t) &= ax_1(t) - bx_1(t)x_2(t) - ex_1(t)^2 \\ \dot{x}_2(t) &= -cx_2(t) + dx_1(t)x_2(t)\end{aligned}\tag{2.8}$$

with parameters a, b, c, d and $e > 0$ Lotka-Volterra model with limited resources.

Similar to (2.6), we can apply the coordinate transformations $x_1 \rightarrow \frac{d}{c}x_1$ and $x_2 \rightarrow \frac{bd}{da-ec}x_2$ to obtain

$$\begin{aligned}\dot{x}_1(t) &= \alpha x_1(t)(1 - x_2(t)) + \beta x_1(t)(1 - x_1(t)) \\ \dot{x}_2(t) &= -cx_2(t)(1 - x_1(t))\end{aligned}\tag{2.9}$$

with $\alpha = a - ec/d$ and $\beta = ec/d$. Here, we have to be careful that positive x_1 and x_2 are mapped on positive values. Since a, b, c, d and $e > 0$ this is the case if and only if $\frac{bd}{da-ec} > 0$, i.e. if $da > ec$.

For $da \leq ec$ one can show that the set of objects $x_2 \rightarrow 0$ for $t \rightarrow \infty$. Here, we want to treat the more interesting case of two coexisting set of objects. Henceforth $da > ec$, which is a necessary condition for the respective setting.

In particular, we obtain the following properties, which are also illustrated in Figure 2.5.

Corollary 2.22 (Equilibria and convergence).

Given the Lotka-Volterra model with limited resources (2.9) the equilibria are given by $x^* = (0, 0)^\top$, $x^{**} = ((\alpha + \beta)/\beta, 0)^\top$ and $x^+ = (1, 1)^\top$. Since only x^+ is an element of $\mathbb{R}^+ \times \mathbb{R}^+$, we obtain

$$Df(x) = \begin{pmatrix} \alpha(1 - x_2) + \beta(1 - 2x_1) & -\alpha x_1 \\ cx_2 & -c(1 - x_1) \end{pmatrix},$$

which gives us

$$Df(x^+) = \begin{pmatrix} -\beta & -\alpha \\ c & 0 \end{pmatrix}$$

and Eigenvalues $\lambda_{1/2} = -\frac{\beta}{2} \pm \sqrt{\frac{\beta^2}{4} - ca}$ rendering x^+ to be exponentially stable.

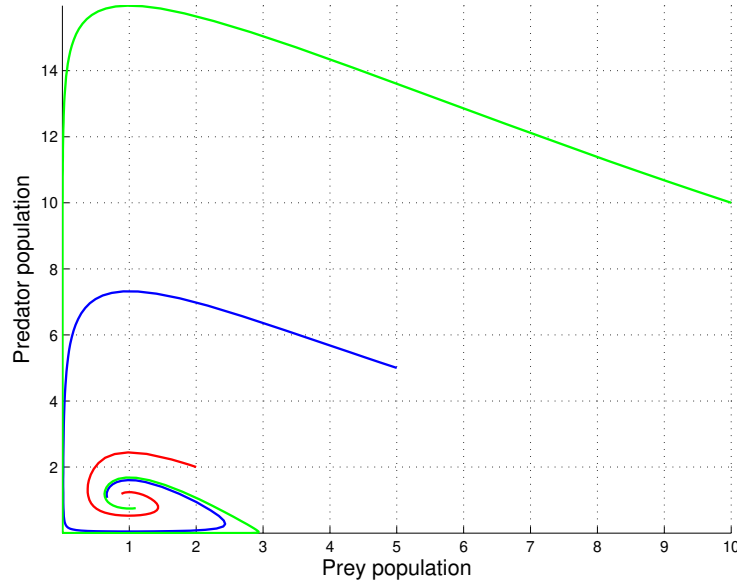


Figure 2.5.: Solutions for the predator–prey model (2.9) with $a = c = 1$ and $\beta = 0.5$

In principle, the model can be extended to n different types of objects x_1 to x_n . If we consider identical model assumptions for all types, where the rate of creation γ depends affine linearly on the other types, we obtain

Definition 2.23 (Generalized Lotka-Volterra model).

We call the system

$$\dot{x}_i(t) = k_i x_i(t) + b_i^{-1} \sum_{j=1}^n a_{ij} x_i(t) x_j(t), \quad i = 1, \dots, n \quad (2.10)$$

with $k_i \neq 0$, $a_{ii} \leq 0$ and $b_i > 0$ *generalized Lotka-Volterra model*.

Note that models (2.4) and (2.8) are special cases of this model.

The special case $a_{ii} = 0$ and $a_{ij} = -a_{ji}$ is called a *Volterra ecology*. In that case, the matrix $A = (a_{ij})$ is anti-symmetric, i.e. $x^\top Ax = 0$ for all $x \in \mathbb{R}^{n_x}$ and $\frac{d}{dt}V(x(t)) = 0$. Hence, we will obtain similar periodic phenomena.

Table 2.2.: Advantages and disadvantages of Lotka-Volterra

Advantage	Disadvantage
✓ Allows multiple species	✗ Requires complex analysis
✓ Allows usage of Lyapunov theory	✗ Hinders property validation
✓ Allows periodic solutions	

CHAPTER 3

MECHANICAL PROCESSES

In this chapter, we consider mechanical processes, which can be modeled using force and velocities. To this end, we first introduce basic translational and rotational models of elements. Then, we utilize Newton's law to develop elementary equations of motion and by that the differential equation itself. The approach itself is constructive and — in principle — allows us to model arbitrarily complex mechanical system at very low mathematical costs. As the approach is impracticable for complex systems, we consider the continuing method of Lagrange.

3.1. d'Alembert Principle

Within this first section, we introduce an approach known as *d'Alembert Principle*. It represents a modularization and combination of mechanical systems. Each of the modules (or elements) is described by a graphical symbol and a respective equation of motion, which, however, not always corresponds to a differential equation.

Here, we start by introducing the modules including the respective equations. For such modules, one distinguishes between two different kinds of motion, translational and rotational. In the following, we will use the denotation given in Table 3.1.

Variable	Meaning	Unit	
m	Mass	kg	[kilogramm]
h	Height	m	[meter]
g	Gravitation	m/s^2	[meter per second square]
E	Energy	$kg\ m^2/s^2$	[Joule]

Table 3.1.: Denomination for technical elements and models

3.1.1. Translational elements

Translational elements are elements of motion, which allow for a movement along a straight line, i.e. a one-dimensional movement. We subdivide these elements in mass, spring and damping ones using the denomination displayed in Table 3.2.

Variable	Meaning	Unit	
y	Location, dilation	m	[meter]
v	Velocity	m/s	[meter per second]
a	Acceleration	m/s^2	[meter per second square]
F	Force	$N = kg\ m/s^2$	[Newton]

Table 3.2.: Denomination for translational models

Mass element

A mass element consists of a mass m (which is constant in time), a force F applied to this mass and the velocity v of the mass (both of which can be time dependent). The symbol for a mass element is depicted in Figure 3.1.

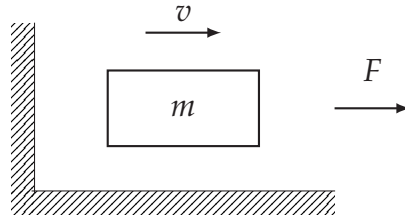


Figure 3.1.: Symbol for a mass element

Utilizing Newton's second law, the differential equation for the mass element is given by

$$F(t) = ma(t) = m\dot{v}(t). \quad (3.1)$$

where force F and velocity v point into the same direction.

If a mass is in motion, then its kinetic energy is given by

$$E_k(t) = \frac{m}{2}v(t)^2.$$

If a mass is caught in a gravity field, then its potential energy is given by

$$E_p(t) = mgh(t).$$

Spring element

The spring (or more generally the elasticity) element is a deformable object, for which the dilation y is a function of the applied force F (which may be time dependent). The symbol for a mass element is given in Figure 3.2.

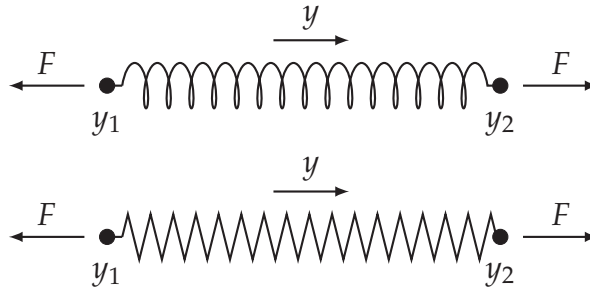


Figure 3.2.: Symbols for a spring element

For the ansatz of a linear model we use Hook's law to describe the spring element. Hence, we have

$$sy(t) = F(t) \quad (3.2)$$

where $y(t) = y_2(t) - y_1(t)$ is the dilation of the spring and $s > 0$ the spring constant. By convention, $y_2(t)$ is the point of action in positive direction, and $y_1(t)$ for negative direction.

For small dilations, this model describes a real life spring sufficiently well. For more realistic models, a nonlinear mapping between $F(t)$ and $y(t)$ is applied, which we will not cover here. Independent from the modeling of this mapping, pure spring elements are an idealization by themselves. In reality, there exists no spring without mass and damper. Note that for $y(t) = 0$, the spring is in a position of rest, hence the dilation can be either positive or negative within this model.

Similar to mass elements, also spring elements can store potential energy. If equation (3.2) is supposed to hold, then this energy is given by

$$E_p(t) = \frac{s}{2}y(t)^2.$$

Damper element

A damper or damping element is a mechanical element, which cannot store energy, but instead converts the received energy into heat and releases the latter. This is referred to as a dissipator. The symbol for a damper element is given in Figure 3.3.

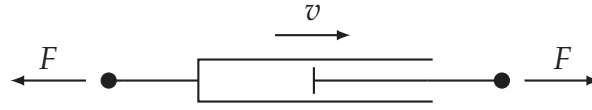


Figure 3.3.: Symbol for a damper element

Again, we consider the linear model given by

$$F(t) = d\dot{v}(t), \quad (3.3)$$

where $v(t)$ is the relative velocity of the body (which corresponds to the piston in the cylinder), $F(t)$ the attacking force and $d > 0$ the damping constant. If a force $F(t)$ is applied, then the velocity $d\dot{v}(t)$ will be reached. The relative velocity $v(t)$ is computed via $v(t) = v_+(t) - v_-(t)$, where $v_+(t)$ denotes the velocity of the terminal point in positive direction, and $v_-(t)$ the velocity of the terminal point in negative direction.

This model is also called *viscosity model* or *viscous friction*. Other models are given by, e.g., *dry friction* or *drag/air resistance*. In the first case, the force $F(t)$ is increasing for slower velocities, in the latter the force quadratically depends on the velocity via $F(t) = d\dot{v}(t)|v(t)|$. Even more complex connections arise in the case of *stiction*, which cannot be modeled by a classical function, but required hysteresis models instead.

The absorbed energy of a damping element at time t is the product $F(t)v(t)$. Hence, in the time interval $[t_0, t_1]$, a damping element absorbs the energy given by the integral over the power, i.e.

$$E_a = \int_{t_0}^{t_1} F(t)v(t)dt.$$

3.1.2. Rotational elements

Analog to translational element, we introduce three elements for rotations. To this end, we use the denomination displayed in Table 3.3.

Variable	Meaning	Unit	
θ	Angle	rad	[radian]
ω	Angular velocity	rad/s	[radian per second]
α	Angular acceleration	rad/s^2	[radian per second square]
τ	Torque	Nm	[Newton meter]
J	Moment of inertia	kgm^2	[kilogramm meter square]

Table 3.3.: Denomination for rotational models

The torque describes the force, which is applied to a rotating body: Consider $F(t) = (F_1(t), F_2(t), 0)$ to be a directed force and a body, which is rotating around the x_3 axis. The force is applied at the body at point $x(t) = (x_1(t), x_2(t), 0)$ as illustrated in Figure 3.4. The vector x can be interpreted

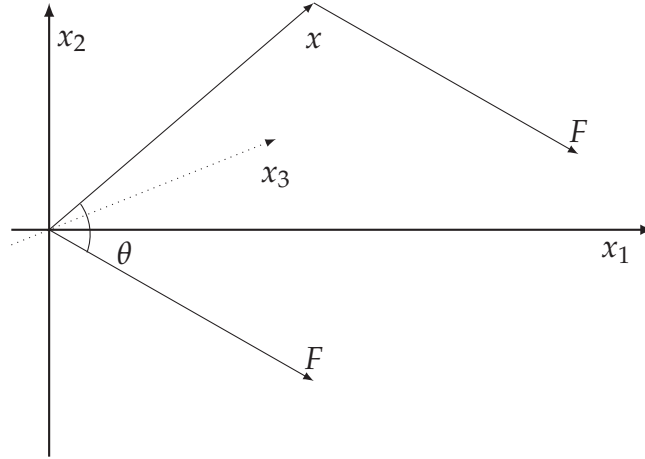


Figure 3.4.: Schematic illustration of torque

as a leverage of the body. The resulting torque is given by

$$\tau(t) = x_1(t)F_2(t) - x_2(t)F_1(t) = \|x(t)\| \|F(t)\| \sin(\theta(t)), \quad (3.4)$$

where $\theta(t)$ is the angle between $x(t)$ and $F(t)$. Again, the sign is important. Positive direction must be chosen such that both expressions in (3.4) coincide.

Note that the force $F(t)$ is now a vector in a coordinate system. In contrast to translational models, the information regarding direction is contained in $F(t)$, hence we can compute contact forces without having to take care of directions.

Mass element

The mass element for rotations consists of a mass, which is rotating around an axis. The respective formula is given by

$$\tau(t) = J\alpha(t) = J\dot{\omega}(t) \quad (3.5)$$

where J represents the moment of inertia, which is given by the mass of the object and its distribution around the rotation axis. Figure 3.5 give the symbol for the rotational mass element.

For a rotating body $B \subset \mathbb{R}^3$ with mass m and density $\rho : B \rightarrow \mathbb{R}_0^+$ we have

$$J = \int_B r(x)^2 \rho(x) dx$$

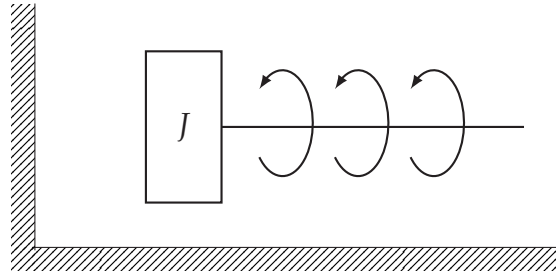


Figure 3.5.: Schematic illustration of rotational mass element

where $r(x)$ is the distance of x to the rotation axis.

In special cases, a closed formula is known. A rotating point mass with mass m and distance r to the rotation axis possesses the moment of inertia

$$J = mr^2.$$

Note that the latter can be generalized to the Parallel Axis Theorem (also known as Steiner's Theorem).

Spring/torsion element

The spring and the following damper element are completely analog to their translational counterparts. Similarly, we consider the linear models only. For the rotational spring element the equation reads

$$s\theta(t) = \tau(t). \quad (3.6)$$

Figure 3.6 gives the respective symbol.

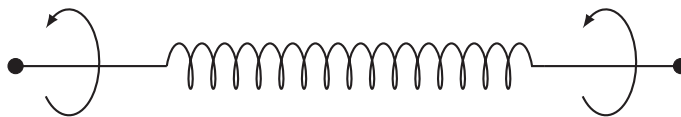


Figure 3.6.: Symbols for a rotational spring element

Damper element

For the damper element, the following equation

$$d\alpha(t) = \tau(t) \quad (3.7)$$

holds and the symbol of the damper element is given in 3.7

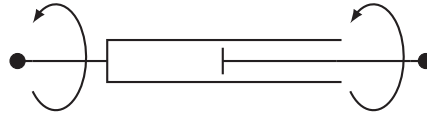


Figure 3.7.: Symbol for a rotational damper element

3.1.3. Combining elements

Based on the previous elements, the Ansatz of the d'Alembert method is to build more complex system by the following procedure:

1. Model the mechanical system using mass, spring and damping elements
2. Prepare the respective equations of motion
3. Formulate the connecting laws / contact forces

The basis for this Ansatz is given by Newton's 3rd Law *actio = reactio*: In each mass, the sum of forces is zero.

Theorem 3.1 (Newton's 3rd law).

Consider a mechanical element. Let F_k for $k = 1, \dots, k_{\max}$ be the internal forces and F be the external force applied to the considered element. Then we have

$$F = \sum_{k=1}^{k_{\max}} F_k.$$

Note that the direction of the force needs to be taken into account using a respective sign. Here, we will exemplary discuss how this procedure works using a pendulum model.

Task 3.2 (Inverted pendulum)

We utilize rotational elements to generate a model of a pendulum and impose the following assumptions:

- The pendulum is a point mass m , which is mounted on a massless rod of length ℓ .
- There is no friction.

Let $x(t) = (x_1(t), x_2(t))^T$ be the endpoint of the pendulum. The rotation axis is located at x_A , and we set $x_A = 0$. As usual, the coordinates x_1, x_2 are increasing rightwards and upwards respectively. A schematic sketch of the model is given in Figure 3.8.

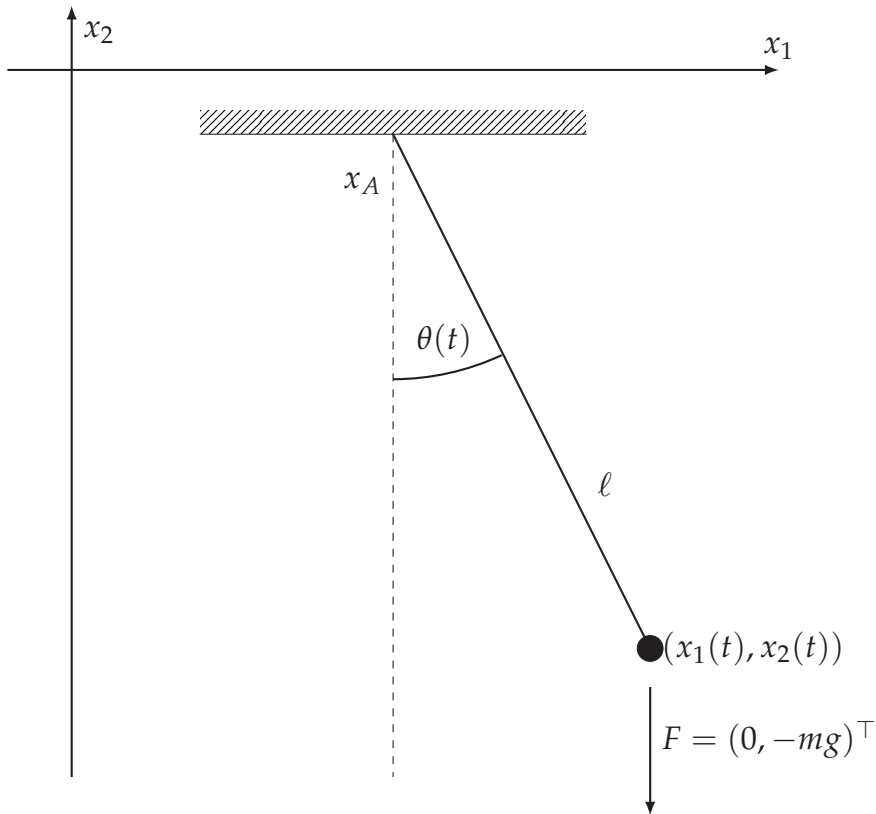


Figure 3.8.: Schematic drawing of a pendulum

Solution to Task 3.2: The point $x(t)$ can be calculated from the length ℓ and the angle $\theta(t)$ via

$$x(t) = (\ell \sin(\theta(t)), -\ell \cos(\theta(t)))^T.$$

Due to earth's gravitation, the force F acting in $x(t)$ is given by $F = (0, -mg)^T$. Utilizing (3.4) we obtain the torque

$$\tau_F(t) = x_1(t) \cdot (-mg) + x_2(t) \cdot 0 = -mgx_1(t) = -mg\ell \sin(\theta(t)).$$

Moreover, for the mass element we obtain from equation (3.5)

$$\tau_J(t) = J\ddot{\theta}(t) = m\ell^2\ddot{\theta}(t).$$

Setting $\tau_F = \tau_J$, we get

$$m\ell^2\ddot{\theta}(t) = -mg\ell \sin(\theta(t)),$$

which gives a second order differential equation. Via $\omega(t) = \dot{\theta}(t)$, we arrive at the system of first order differential equations

$$\begin{aligned}\dot{\theta}(t) &= \omega(t) \\ \dot{\omega}(t) &= -\frac{g}{\ell} \sin(\theta(t)).\end{aligned}$$

Combining properties of d'Alembert's method, we see the contents of Table 3.4.

Table 3.4.: Advantages and disadvantages of d'Alembert's method

Advantage	Disadvantage
✓ Allows to connect base elements	✗ Requires connecting points
✓ Based on simple principle	✗ Complex for large problems
✓ Easily accessible	✗ No direct meaning of energy

3.2. Lagrangian formalism

Within the last section we discussed a method to combine basic translational and rotational models with their forces. For large systems, this procedure is rather complex. The reason lies in the number of connection laws and contact forces for many points, each resulting in a single equation. This leads to large equation systems, which are difficult to solve.

An alternative is the so called *energy based* method using *Lagrange–Equations*. The idea of the Lagrange–Equations utilizes the energy of a system. We restrict ourselves to the case of a system with n points of mass m_i at locations $r_i = (x_i, y_i, z_i)^\top$, $i = 1, \dots, n$. The kinetic energy of this

system is given by

$$E_k = \sum_{i=1}^n \frac{m_i}{2} \|v_i\|^2.$$

The mechanical structure with its J connections induces constraints, which can be formalized as follows:

Definition 3.3 (Constraints).

Consider a system with n points of mass m_i at locations $r_i = (x_i, y_i, z_i)^\top, i = 1, \dots, n$. Then we denote the constraints of the system by

$$C_n(r_1, \dots, r_n, t) = 0 \quad \forall n = 1, \dots, J, \quad (3.8)$$

where $r_i = (x_i, y_i, z_i)^\top \in \mathbb{R}^3$ marks the positions of the points of mass.

Task 3.4 (Pendulum constraints)

Consider the pendulum from Task 3.2 fixed at the origin with point mass m at point $r(t) = (x(t), y(t), z(t))^\top$ of length ℓ , which is swinging in the $x - y$ plain. Formulate the constraints (3.8) for this system.

Solution to Task 3.4: All possible positions of $r(t)$ are the given by

$$C_1 = \|r\|^2 - \ell^2 \quad \text{and} \quad C_2(r) = z.$$

Let us now assume that we can parameterize the *manifold of compatible configurations*, then we get the following:

Definition 3.5 (Generalized coordinates).

Consider a system with constraints (3.8). Then we call the set of coordinates $q(t) = (q_1(t), \dots, q(t)) \in Q \subset \mathbb{R}^{n_q}$ *generalized coordinates* if they satisfy

$$\begin{aligned} M &= \left\{ (r_1(q(t), t), \dots, r_n(q(t), t))^\top \mid q(t) \in Q \right\} \\ &= \left\{ (r_1, \dots, r_n)^\top \mid C_j(r_1, \dots, r_n, t) = 0 \ \forall j = 1, \dots, J \right\} \end{aligned}$$

for continuously differentiable functions $r_i(q, t)$ and

$$\frac{\partial r}{\partial q_k}(q(t), t), \quad k = 1, \dots, n_q \text{ are linearly independent.}$$

Task 3.6 (Pendulum generalized coordinates)

Design a set of generalized coordinates for the pendulum example from Task 3.2

Solution to Task 3.6: For the pendulum we have

$$r(q(t)) = \begin{pmatrix} \ell \sin(q(t)) \\ -\ell \cos(q(t)) \\ 0 \end{pmatrix}$$

with $q(t) = q_1(t) \in Q = (-\varepsilon, 2\pi) \subset \mathbb{R}$ for arbitrary $\varepsilon > 0$. Note that q describes the angle of the pendulum, which is denoted by θ in the previous section.

Now we can describe our system using the generalized coordinates $q(t)$. Via the chain rule, we can also express the velocity in terms of $q(t)$. We obtain

Definition 3.7 (Generalized velocities).

Consider a system with generalized coordinates $q(t) = (q_1(t), \dots, q_{n_q}(t)) \in Q \subset \mathbb{R}^{n_q}$. Then we call the set $\dot{q}_1, \dots, \dot{q}_{n_q}$ given via

$$v_i(t) = \frac{d}{dt}r_i(q(t), t) = \sum_{j=1}^J \frac{\partial r_i}{\partial q_j}(q(t), t)\dot{q}_j(t) + \frac{\partial r_i}{\partial t}(q(t), t), \quad i = 1, \dots, n. \quad (3.9)$$

generalized velocities.

Note that due to linear independence of the partial derivatives, this equation system (3.9) can be solved for $\dot{q}(t)$.

Task 3.8 (Pendulum generalized velocities)

Derive the generalized velocities 3.9 for the pendulum example of Task 3.2.

Solution to Task 3.8: For the pendulum we have

$$v(t) = \begin{pmatrix} \ell \cos(q(t)) \\ \ell \sin(q(t)) \\ 0 \end{pmatrix} \dot{q}(t).$$

Now, we can write the kinetic energy using q and \dot{q} via

Definition 3.9 (Generalized kinetic energy).

Consider a system with generalized coordinates and velocities. Then we call

$$E_k = \sum_{i=1}^n \frac{m_i}{2} \|v_i\|^2 = \sum_{i=1}^n \frac{m_i}{2} \left\| \sum_{j=1}^J \frac{\partial r_i}{\partial q_j}(q(t), t) \dot{q}_j(t) + \frac{\partial r_i}{\partial t}(q(t), t) \right\|^2 =: \mathcal{T}(q(t), \dot{q}(t), t)$$

kinetic energy, which is also denoted by $\mathcal{T}(q(t), \dot{q}(t), t)$.

Task 3.10 (Pendulum generalized kinetic energy)

Given the pendulum example from Task 3.2 compute the generalized kinetic energy.

Solution to Task 3.10: For the pendulum we have

$$\mathcal{T}(q(t), \dot{q}(t), t) = \frac{m}{2} \ell^2 \dot{q}(t)^2$$

as generalized kinetic energy.

For forces $F_i(t) \in \mathbb{R}^3$, $i = 1, \dots, n$, which are applied at the i th point of mass, we define the so called *generalized forces* via

Definition 3.11 (Generalized forces).

Consider a system with generalized coordinates. Then we call the set

$$f_k(t) = \sum_{i=1}^n \left\langle F_i(t), \frac{\partial r_i}{\partial q_k}(q(t), t) \right\rangle, \quad k = 1, \dots, n_q.$$

generalized forces. Furthermore, we call a mechanical system *conservative*, if there exists a function $\mathcal{W}(r(q(t), t), t)$ such that

$$F_i(t) = -\frac{\partial \mathcal{W}}{\partial r_i}(r(q(t), t), t) =: -\nabla_i \mathcal{W}(r_1(q(t), t), \dots, r_n(q(t), t), t)$$

holds.

The definition of generalized forces directly leads to the potential energy of a system.

Definition 3.12 (Generalized potential energy).

Consider a system with generalized forces. We call \mathcal{W} *generalized potential energy* if $f(t) = -\nabla_q \mathcal{W}(q(t), t)$ holds for $\mathcal{W}(q(t), t) = \mathcal{W}(r(q(t), t), t)$.

Since the generalized potential energy is defined via derivatives, one typically adds a suitable constant to arrive at $\min_q \mathcal{W}(q(t), t) = 0$.

Task 3.13 (Pendulum generalized potential energy)

Compute the generalized forces / potential energy for the pendulum from Task 3.2.

Solution to Task 3.13: Utilizing the pendulum example without friction, the force $F(t) = (0, -mg, 0)^\top$ applies to the pendulum, which can be written as $f(t) = -\nabla_q \mathcal{W}(q(t), t)$ with $\mathcal{W}(q(t), t) = mgy(t)$. Inserting $r(q(t)) = (\ell \sin(q(t)), -\ell \cos(q(t)), 0)^\top$, we have $\mathcal{W}(q(t), t) = -mg\ell \cos(q(t))$. To satisfy $\min_q \mathcal{W}(q(t), t) = 0$, we add $mg\ell$ to the expression and obtain

$$\mathcal{W}(q(t), t) = -mg\ell \cos(q(t)) + mg\ell$$

as generalized potential energy.

Having defined the notation above, we are now ready to define the Lagrangian:

Definition 3.14 (Lagrangian).

Consider a conservative mechanical system. Then we call the function

$$\mathcal{L}(q(t), \dot{q}(t), t) = \mathcal{T}(q(t), \dot{q}(t), t) - \mathcal{W}(q(t), t) \quad (3.10)$$

the Lagrangian of the system.

Utilizing the Lagrangian, we can derive the equations of motion of the system:

Theorem 3.15 (Lagrangian equation).

Consider a conservative mechanical system. Then the so called Lagrangian Equation

$$\frac{d}{dt} \left(\frac{\partial \mathcal{L}}{\partial \dot{q}_k}(q(t), \dot{q}(t), t) \right) - \frac{\partial \mathcal{L}}{\partial q_k}(q(t), \dot{q}(t), t) = 0, \quad k = 1, \dots, n_q \quad (3.11)$$

holds.

Note that this equation can be obtained from the physical condition that the functional

$$\mathcal{I}(q(t)) = \int_{t_0}^{t_1} \mathcal{L}(q(t), \dot{q}(t), t) dt$$

is minimal along solutions q . Setting $g(\alpha) = \mathcal{I}(q + \alpha z)$ for an arbitrary differentiable function z with $z(t_0) = z(t_1) = 0$, then we have $\dot{g}(0) = 0$. After some computations we obtain

$$\dot{g}(0) = \sum_{k=1}^{n_q} \int_{t_0}^{t_1} \left(\frac{d}{dt} \left(\frac{\partial \mathcal{L}}{\partial \dot{q}_k}(q(t), \dot{q}(t), t) \right) - \frac{\partial \mathcal{L}}{\partial q_k}(q(t), \dot{q}(t), t) \right) z(t) dt$$

revealing (3.11).

Task 3.16 (Pendulum Lagrangian)

Considering the pendulum from Task 3.2, derive the Lagrangian equation of motion.

Solution to Task 3.16: Considering the pendulum without friction, we obtain

$$\mathcal{L}(q(t), \dot{q}(t), t) = \frac{m}{2} \ell^2 \dot{q}(t)^2 + mg\ell \cos(q(t)) - mg\ell.$$

Hence, we have

$$\begin{aligned} \frac{\partial \mathcal{L}}{\partial \dot{q}}(q(t), \dot{q}(t), t) &= m\ell^2 \dot{q}(t) \\ \frac{\partial \mathcal{L}}{\partial q}(q(t), \dot{q}(t), t) &= -mg\ell \sin(q(t)). \end{aligned}$$

Hence, we obtain the equations of motion via (3.11)

$$\begin{aligned} 0 &= \frac{d}{dt} \left(m\ell^2 \dot{q}(t) \right) + mg\ell \sin(q(t)) \\ &= m\ell^2 \ddot{q}(t) + mg\ell \sin(q(t)). \end{aligned}$$

Since $\ell > 0$ and $m > 0$, the latter simplifies to

$$0 = \ell \ddot{q}(t) + g \sin(q(t)),$$

which corresponds to our earlier results with $q = \theta$.

Remark 3.17

The Lagrangian approach we presented here is given for conservative systems, i.e. systems without loss of energy, e.g., via friction. To integrate such effects gives us a so called dissipative system. Within the modeling, a dissipation rate needs to be defined and translated into a generalized friction force. Then, we can add this force to the right hand side of the Lagrangian Equation (3.11) and solve the latter.

Combining properties of the Lagrangian approach, we see the contents of Table 3.5.

Table 3.5.: Advantages and disadvantages of the Lagrangian/Hamiltonian approach

Advantage	Disadvantage
✓ Based on energy	✗ Requires generalized coordinates
✓ Simple derivation of motion	✗ Requires indepth theory
✓ Possible automation of approach	✗ Reveals second order system

CHAPTER 4

ELECTRICAL PROCESSES

In contrast to mechanical processes, electrical ones do not work on force and velocities, but instead on voltage and currents. To model such processes, we first introduce the base components of generation, transmission, transformation, storage, and utilization of electrical energy. Thereafter, we introduce the concept of bond graphs. The latter can also be seen as an overarching method to model both electrical networks as well as mechanical applications.

4.1. Network elements

Bond graphs are a graphical representation used to model the dynamic behavior of physical systems. They are particularly useful for systems involving multiple energy domains, such as mechanical, electrical, hydraulic, and thermal systems. Bond graphs employ a set of standardized symbols to represent the components and interactions within these systems, focusing on the flow of energy between elements. The basis of the latter is a so called *graph* or *network*.

Definition 4.1 (Network/graph).

Consider a set of $\mathcal{V} = \{v_1, \dots, v_{n_{\mathcal{V}}}\}$ where $n_{\mathcal{V}} \in \mathbb{N}$ is the maximal entry of \mathcal{V} . Moreover, suppose $\mathcal{E} = \mathcal{V} \times \mathcal{V}$ where $n_{\mathcal{E}} \in \mathbb{N}$ is the maximal entry of \mathcal{E} . Then we call \mathcal{V} the *set of vertexes*, \mathcal{E} the *set of edges* connecting the vertexes, and $\mathcal{N} = (\mathcal{V}, \mathcal{E})$ a *network* or *graph*.

The network/graph is called *directed* if the set of edges is defined via start and ending points, otherwise it is called *undirected*.

Task 4.2

Consider a system given by $\mathcal{V} = \{A, B, C, D, E, F\}$, which is complete, i.e. for each pair of distinct vertexes there exists an edge connecting the latter. Draw the respective network.

Solution to Task 4.2: Within a complete network for each pair of vertexes there exists an edge. The network is displayed in Figure 4.1.

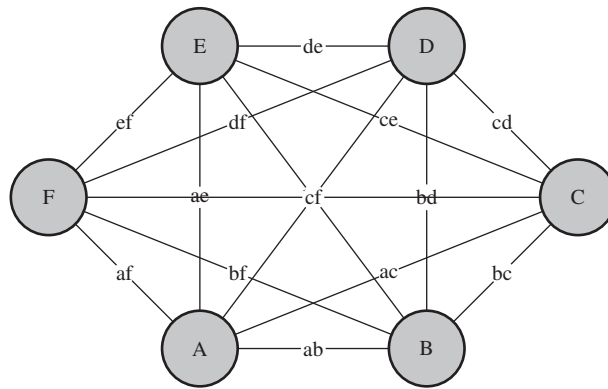


Figure 4.1.: Graph of network from Task 4.2

Within bond graphs, a certain denomination is used for both vertexes, edges and connectivity.

Definition 4.3 (Undirected bond graph).

An *undirected bond graph* is an undirected graph where

- vertices called *nodes* denote subsystems, components or basic elements, and
- edges called (*power*) *bonds* represent the instantaneous energy transfer between nodes,
- connection points of a node are called *power ports*, and
- nodes are called *multiport* if they exhibit more than one power port.

Task 4.4

Consider the bond graph in Figure 4.2. Identify the nodes and their multiport properties.

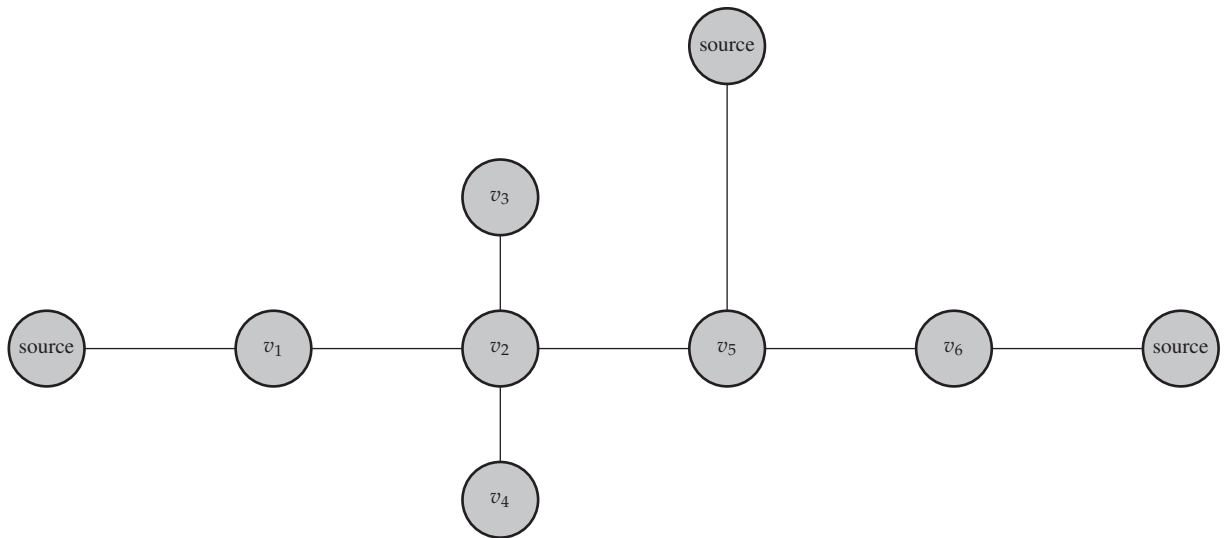


Figure 4.2.: Example of a bond graph

Solution to Task 4.4: The set of nodes is given by $\mathcal{V} = \{v_1, v_2, v_3, v_4, v_5, v_6\}$ where v_3, v_4 are one-ports, v_1, v_6 are two-ports while v_5 is a three-port and v_2 is a four-port.

The notion of power bonds and power ports indicates that bond graphs deal with the power transfer within graphs. In general, power p equals the product of two physical quantities, the so called *flow* f and *effort* e , i.e.

$$p = e \times f.$$

By convention, in the horizon case of a power bond effort is always written on top and flow below a power bond. In the vertical case, effort is written on the left and flow on the right, cf. Figure 4.3.



Figure 4.3.: Convention for annotation of power bonds

Remark 4.5

As a consequence of using flow and effort, a bond graph is an energy based representation

whereas a block diagram is a signal based representation. As a consequence, bond graphs can be applied in multiple domains (simultaneously).

Within the literature, two analogies for effort and flow between disciplines are known (but contradict one another). Here, we apply the so called *direct analogy* stating the possibilities outlined in Table 4.1.

Table 4.1.: Direct analogy for power

	Effort	Flow	Momentum	Displacement
Translational mechanics	force	velocity	momentum	displacement
Rotational mechanics	moment	velocity	momentum	angle
Electromagnetics	voltage	current	linkage flux	charge
	force	flux rate		flux
Hydraulics	pressure	volume flow	momentum	volume
Thermodynamics	temperature	entropy flow		entropy
Chemics	potential	molar flow		mass

While power exchange can be represented by undirected edges, evaluation of models in nodes require signs for the latter. In bond graphs, the following is used:

Definition 4.6 (Directed bond graph).

Suppose a bond graph to be given. Then it is called *directed* if a half arrow is used for each power bond to indicate the positive reference direction of the flow f across the bond.

By convention, the half arrow is added to the side where the flow variable is annotated, cf. Figure 4.4.

4.2. Bond graph junctions

Physical processes suggest the introduction of classes for basic multiport elements



Figure 4.4.: Convention for annotation of power bonds

- sources and sinks
- storages
- dissipators
- power couplers and transducers, and
- power nodes for distribution.

Definition 4.7 (Junction structure).

Suppose a bond graph to be given. If the nodes transfer or distribute power instantaneously only, the it is called *junction structure*.

In particular, we consider the following junctions characterizing equal effort (0-junction) or equal flow (1-junction). More formally:

Definition 4.8 (0-junction).

A *0-junction* is a multiport characterized by

$$e_1 = e_2 = \dots = e_n \quad (4.1)$$

$$\sum_{k=1}^n f_k = 0 \quad (4.2)$$

where n is the cardinality of the multiport.

The latter corresponds to Kirchhoff's current law in electrics or the description of links in mechanics.

Definition 4.9 (1-junction).

A multiport is called *1-junction* if it satisfies

$$\sum_{k=1}^n e_k = 0 \quad (4.3)$$

$$f_1 = f_2 = \dots = f_n \quad (4.4)$$

where n is the cardinality of the multiport.

Among others, the 1-junction corresponds to Kirchhoff's voltage law or d'Alembert's principle on velocities. Based on the power connections established by the power ports, information can be fed to other nodes, e.g. into a block diagram for measurement or control. Such ports do not handle power but signals instead, which gives us:

Definition 4.10 (Signal port).

A port is called *signal port* if it provides information regarding the flows and efforts connected to a node.

These signal ports allow us to connect the information structure of a block diagram to the power structure of the bond graph. In diagrams, such signals are indicated by full headed arrows, cf. Figures 4.5a and 4.5b.

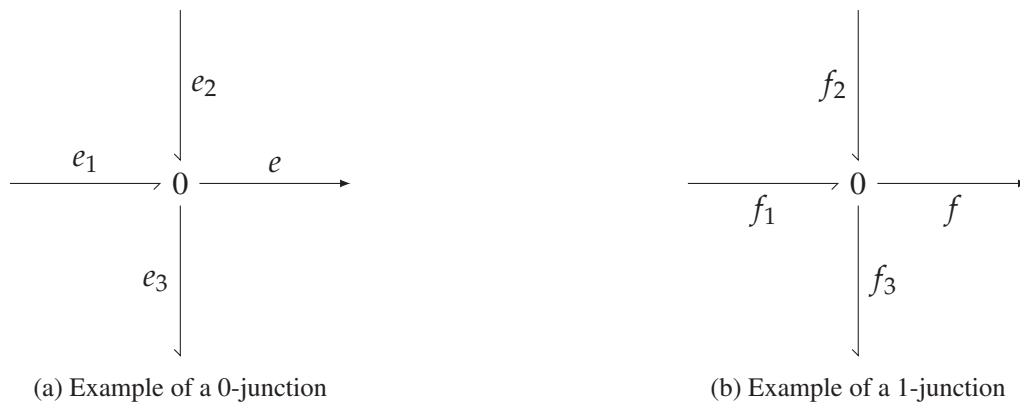


Figure 4.5.: Example of 0- and 1-junctions

We can also combine 0- and 1-junctions to a more complex structure. Between such elements, the bonds can be used to define an internal structure.

Definition 4.11 (Internal bond).

A bond is called *internal bond* if it connects a 0- or 1-junction to another 0- or 1-junction.

Note that it does not make sense to combine two 0-junction or two 1-junctions as they can simply be summarized into a single node. Hence, internal bonds need only be considered between 0- and 1-junctions. Together, an internal structure arises:

Definition 4.12 (Simple junction structure).

A bond graph is called to be of *simple junction structure* if each node is either a 0- or a 1-junction.

The simple junction structure is also called Kirchhoff structure. Everything outside of such a simple junction structure is called external, also the connecting bond.

Definition 4.13 (External bond).

If a bond connects a 0- or 1-junction to a power port of an element that does not belong to the simple junction structure, then it is called *external bond*.

4.3. Bond graph modeling

On the outside of simple junction structures, different elements can be used.

Definition 4.14 (Two-port transformer).

A node is called *two-port transformer* if it satisfies the conditions

$$e_1(t) = m(\cdot) \times e_2(t) \quad (4.5)$$

$$m(\cdot) \times f_1(t) = f_2(t) \quad (4.6)$$

where $m(\cdot) \in \mathbb{R}^+$ is either a constant, a function of time or a function of another power variable.

Note that by definition, the modulus may arise from both efforts and flows, yet with inverted impact. In diagrams, we apply the notation shown in Figure 4.6, which also indicates whether $m(\cdot)$ enters the node as a constant or function.

Within Figure 4.6, MTF stands for modulated transformer.



Figure 4.6.: Representation of a two-port transformer

Task 4.15

Consider a mechanical gear with radii r_1 and r_2 . Define the two-port transformer describing the gear.

Solution to Task 4.15: The tangial velocity define the flow equation

$$f_1 = r_1 \times \omega_1 = r_2 \times \omega_2 = f_2$$

where ω_1, ω_2 are the rotation speeds of gears. Hence, for the forces we obtain

$$r_2 \cdot e_1 = r_1 \times e_2$$

where the efforts e_1, e_2 correspond to the moments of the gears.

For a transformer, we saw that inputs are transformed to the same outputs, i.e. flows to flows and efforts to efforts. Conversely, a gyrator can be applied.

Definition 4.16 (Two-port gyrator).

We call a node *two-port gyrator* if it satisfied the conditions

$$e_1 = r \times f_2 \tag{4.7}$$

$$e_2 = r \times f_1 \tag{4.8}$$

with $r \in \mathbb{R}^+$ representing the gyrator ratio being either a constant, a function of time or a function of another power variable.

Similar to the transformer, the gyrator is depicted as given in Figure 4.7 where again MGY stands for modulated gyrator.



Figure 4.7.: Representation of a two-port gyrator

Task 4.17

Consider an electrical coil on a magnetic core with n turns satisfying Faraday's Law

$$u = n \times \frac{d\Phi}{dt}$$

for flux Φ and voltage u . Identify the magnetomotive force and draw the two-port gyrator.

Solution to Task 4.17: Due to power balance with current i

$$u \times i = V \times \frac{d\Phi}{dt} \quad (4.9)$$

we obtain

$$V = n \times i$$

and the gyrator in Figure 4.8.

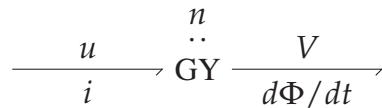


Figure 4.8.: Two-port gyrator according to Task 4.17

Combined, these elements give rise to the so called *general junction structure*.

Definition 4.18 (General junction structure).

We call a bond graph to be of *general junction structure* if each node is either a 0- or 1-junction or an (M)TF or (M)GY.

A special case of the latter is the so called *weighted junction structure*, which does not interchange outputs but instead works like an amplifier/transformator.

Definition 4.19 (Weighted junction structure).

A bond graph with only 0-, 1- or (M)TF nodes is called to be of *weighted junction structure*.

Within the scope of the lecture, we consider everything outside a general junction structure to be sources and sinks or simply disturbances. In terms of bond graphs, we define

Definition 4.20 (Environmental elements).

Nodes are called *environmental elements* if they do not belong to a general junction structure.

Among these elements are storages and resistors, which we consider next.

Definition 4.21 (1-port C energy storage).

A node is called *1-port C energy storage* if there exists a bijective function $\Phi_C : \mathbb{R} \rightarrow \mathbb{R}$ such that the node satisfies

$$q(t) = \Phi_C(e(t)) \quad (4.10)$$

for the effort $e : \mathbb{R} \rightarrow \mathbb{R}$ and generalized displacement $q : \mathbb{R} \rightarrow \mathbb{R}$ with time $t \in \mathbb{R}^+$.

Definition 4.22 (1-port I energy storage).

A node is called *1-port I energy storage* if there exists a bijective function $\Phi_L : \mathbb{R} \rightarrow \mathbb{R}$ such that the node satisfies

$$p(t) = \Phi_L(f(t)) \quad (4.11)$$

for the flow $f : \mathbb{R} \rightarrow \mathbb{R}$ and generalized momentum $p : \mathbb{R} \rightarrow \mathbb{R}$ with time $t \in \mathbb{R}^+$.

Table 4.2.: C and I energy storages

	C storage	I storage
Translational mechanics	Spring	Rigid body
Rotational mechanics	Torsion spring	Flywheel
Continued on next page		

Table 4.2 – continued from previous page

	C storage	I storage
Electromagnetics	Capacitor	Coil
	Ferromagnetic	-
Hydraulics	Fluid compressibility	Fluid inertia
Thermodynamics	Lump of material	-

To illustrate such a network and its translation to a bond graph, we consider the following simple example.

Definition 4.23 (1-port resistor).

We call an node *1-port resistor* if there exists one of the bijective functions $\Phi_R : \mathbb{R} \rightarrow \mathbb{R}$ or $\Phi_G : \mathbb{R} \rightarrow \mathbb{R}$ such that the node satisfies one of the conditions

$$e(t) = \Phi_R(f(t)) \quad (4.12)$$

$$f(t) = \Phi_G(e(t)) \quad (4.13)$$

for effort $e : \mathbb{R} \rightarrow \mathbb{R}$ and flow $f : \mathbb{R} \rightarrow \mathbb{R}$ with time $t \in \mathbb{R}^+$.

Task 4.24

Consider the mechatronic system shown in Figure 4.9. Derive the bond graph modeling the system and the respective equations.

Solution to Task 4.24: The bond graph is given in Figure 4.10. We directly obtain the following:

- For the modulated gyrator we have

$$\tau = \Phi \times i_A$$

$$u_A = \Phi \times \omega.$$

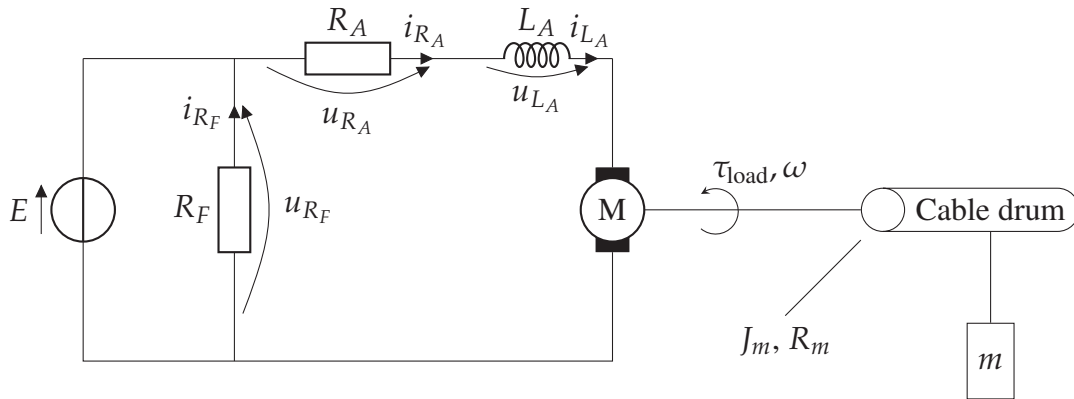


Figure 4.9.: Schematic of a DC machine moving a (hoisting) drum roll

where i_A , u_A are the current and voltage on the inner loop and τ is the torque of the motor with its angular velocity ω .

- For the left 0-junction E represents the voltage supplied by the source. We obtain the current running over the source is given by

$$i = i_{R_A} + i_{R_F}$$

with $i_A = i_{L_A} = i_{R_A}$.

- Continuing from the left, the 1-junction reveals

$$-E + u_{R_A} + u_{L_A} + u_A = 0.$$

- Similarly, the upper 1-junction reveals that the voltage of the source E is given by u_{R_F} .
- Last, the right 1-junction supplies the sum of all torques

$$\tau + \tau_{\text{load}} - \tau_{R_m} - J_m \dot{\omega} = 0.$$

Combined, we can summarize the advantages and disadvantages of Bond graphs as given in Table 4.3.

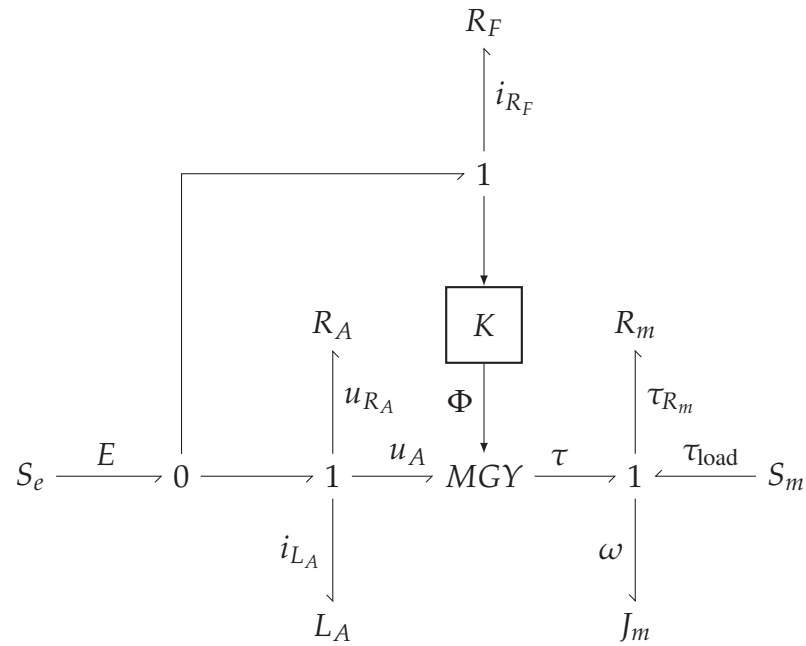


Figure 4.10.: Bond graph of DC driven drum roll

Table 4.3.: Advantages and disadvantages of bond graphs

Advantage	Disadvantage
✓ Allows multiple connections	✗ Requires network structure
✓ Allows power and information bonds	✗ Requires unification
✓ Unifies modeling	

CHAPTER 5

STOCHASTIC PROCESSES

Financial processes are a rather young field of research, which in contrast to the topics considered so far is almost purely stochastic.

Here, we will focus on safeguards for unwanted future development, e.g. in development of exchange rates, and particularly discuss the simplest form known as the *European Option*. Our aim is to compute the value of such a derivative at a given time instant. Since the value depends on the unknown future development of the stocks and rates, we require a respective model of these.

To this end, we apply *stochastic* differential equations, cf. Definition 1.13. For each initial condition, such equations exhibit a number of possible solutions, which depend on chance. The idea of these stochastic differential equations is to approximate possible future developments such that known statistical values from past data (such as expected value or variance) are best modeled. Within this chapter, we first derive a respective model for the most simple task in finance, the assessment of options, and derive models for the stock development. Last, we show a practical method for computing prices of options.

5.1. Options

An option is a contract, which provides the holder with the possibility (but not the obligation) to sell or buy a share at a future time instant for a fixed price. The price is referred to as *strike price*, the selling option is also called a *put* and the buying option is called a *call*.

Here, we consider the *European option*. The difference to other options is that the strike time is apriori fixed. The task now is the following:

What is the value of the option itself at time before the strike time?

Definition 5.1 (Terms of options).

Given an option, we call T *strike time* and denote the (known) *base value* of a share at a certain time $t \in [0, T]$ by S . We denote the *desired value* of the option by $V(t, S)$. Furthermore, let K the fixed *strike price*.

Now, we directly obtain

Corollary 5.2 (Value of an option).

The value of a call option is given by

$$V(T, S) = \begin{cases} S(T) - K & \text{if } S(T) > K \\ 0 & \text{else} \end{cases} = \max\{S(T) - K, 0\} =: (S(T) - K)^+.$$

and of a put option by

$$V(T, S) = \begin{cases} K - S(T) & \text{if } K > S(T) \\ 0 & \text{else} \end{cases} = \max\{K - S(T), 0\} =: (K - S(T))^+.$$

To compute the value of the option at any time $t < T$ we require

- (1) a rule for computing $V(t, S)$ from $V(T, S(T))$ if $S(T)$ is known, and
- (2) an estimate of the base value $S(T)$ at time T depending on the base value S at time t .

If (1) and (2) are available, then we can estimate $V(T, S(T))$ via (2) and apply the rule (1) to this estimate.

We focus on the first requirement first. To this end, we suppose the following to hold:

Assumption 5.3 (Arbitrage freeness)

Given a financial market, no benefit from a risk free fund can be drawn. For a risk free fund, we call $r > 0$ *interest rate* modeling its development.

The respective postulate assumes that if a product is traded at two markets at different prices, then the prices would converge immediately, rendering arbitrage to be impossible. Although this doesn't hold in practice, it is an accepted assumption.

Definition 5.4 (Risk free payoff).

Given a risk free interest rate $r > 0$ and an option $V(\cdot, S)$, then the *payoff* is given by

$$B(T) = \exp^{r(T-t)} V(t, S(t)) \quad (5.1)$$

for any $t < T$.

If we consider the value $V(T, S(T))$ to be known and if $V(t, S(t)) > \exp^{-r(T-t)} V(T, S(T))$, then we could sell the option immediately and invest the payoff risk free. Hence, we obtain

$$B(T) = \exp^{r(T-t)} V(t, S(t)) > V(T, S(T))$$

and our risk free profit is given by $B(T) - V(T, S(T)) > 0$. Vice versa, if $V(t, S(t)) < \exp^{-r(T-t)} V(T, S(T))$, then we could buy that option for $B(t) = V(t, S(t))$ and at strike time get the return

$$B(T) = V(T, S(T)) > \exp^{r(T-t)} V(t, S(t)).$$

Now the risk free profit is given by $B(T) - \exp^{r(T-t)} V(t, S(t)) > 0$. Since the postulate of no-arbitrage bounds from Assumption 5.3 excludes risk free profits, the following holds:

Theorem 5.5 (Value of option).

Suppose Assumption 5.3 to hold and an option $V(\cdot, S)$ to be given. Then the value of the option is given by

$$V(t, S(t)) = \exp^{-r(T-t)} V(T, S(T)).$$

Focusing on the second requirement, we model the typical stock development using a stochastic differential equation of form (1.11)

$$\dot{x}(t) = a(t, x(t)) + b(t, x(t))X(t, \cdot)$$

and set $S(T) = x(T; t, S(t))$. Note that $S(T)$ is not a fixed value but a random variable. The value of $V(T, S(T))$ can be estimated via the expected value $E(V(T, x(T; t, S(t))))$. Hence, we have

Corollary 5.6 (Value of option).

Given an option $V(\cdot, S)$ we suppose that Assumption 5.3 holds. Then the value of the option is given by

$$V(t, S(t)) = \exp^{-r(T-t)} \mathbb{E} (V(T, x(T; t, S(t)))) , \quad (5.2)$$

for any $t \leq T$.

In order to retrieve a dynamics such as (1.11), simplest stochastic differential equation model satisfying these requirements is given by

Definition 5.7 (Geometric Brownian motion).

The solutions of the stochastic differential equation

$$dx(t) = \mu x(t)dt + \sigma x(t)dW_t. \quad (5.3)$$

are called *geometric Brownian motion*

The minimal requirements for modeling a system (5.3) are the parameters *trend* $\mu \in \mathbb{R}$ and the *spreading* $\sigma > 0$. The first parameter μ gives the general direction of the stock development, either up, down or leveling, while the second parameter σ corresponds to the variance/jitter of the stock development around the general direction. In finance, the parameters μ and σ are also termed *rate of return* and *volatility*.

Due and despite its simplicity, the model (5.3) is the basis of many applications regarding the modeling of stocks. In particular, we can show

Theorem 5.8 (Uniqueness Brownian motion).

Consider the stochastic differential equation (5.3), then its unique solution is given by

$$x(t; t_0, x_0) = x_0 \exp^{(\mu - \frac{1}{2}\sigma^2)t + \sigma W(t)}. \quad (5.4)$$

Remark 5.9

For $\sigma = 0$ we reobtain the solution of the linear differential equation $\dot{x}(t) = \mu x(t)$ and its solution $x(t; t_0, x_0) = x_0 \exp^{\mu(t-t_0)}$.

Based on the latter, we can conclude

Corollary 5.10 (Expected value and variance).

For the Brownian motion we can set $e(t) = E(x(t; t_0, x_0))$ and obtain

$$\dot{e}(t) = \mu e(t), \quad e(0) = E(x_0) = x_0 \quad \text{for} \quad (5.5)$$

revealing

$$E(x(t; t_0, x_0)) = x_0 \exp^{\mu(t-t_0)}. \quad (5.6)$$

Moreover, we obtain

$$\sigma^2(x(t; t_0, x_0)) = x_0^2 \exp^{2\mu(t-t_0)} (\exp^{\sigma^2(t-t_0)} - 1). \quad (5.7)$$

The parameters trend μ and spreading σ are typically estimated using past values. This shows, that this type of model is not entirely suited for generating prediction of stock developments. For risk neutral assessment, we set $\mu = r$.

5.2. Monte–Carlo method

Having characterized the Brownian motion as a solution to the stochastic differential equation (5.3), we now use the Wiener process to describe the derivative of the random variable.

Since the Wiener process is a stochastic function, the solutions computed based on a Wiener process are again stochastic functions. Hence, each state x_i in $x = (x_1, x_2, \dots, x_N)^\top$ is a real valued stochastic process connected to one Wiener process $W(t, \omega)$. To mark this connection, for any given path $W(t, \omega)$ we denote the solution by $x(t; t_0, x_0, \omega)$.

The Monte–Carlo method is a direct and very versatile method to compute the expected value of complex expressions. Here, we apply it to compute the expected value $E(V(T, x(T; t, S(t))))$ to assess the value of an option $V(t, S(t))$ via (5.2). Similar to the name giving casino in Monaco, the Monte–Carlo method utilizes a vast number of random experiments. Instead of hoping for a prize, we calculate an estimate of the expected value based on the results of the random experiments. The random experiments themselves are performed by computers according to the following algorithm, and the solution is therefore a numerical and not an analytic one.

Algorithm 5.11 (Monte–Carlo Method)

Given a stochastic differential equation (5.3), an initial time t , a strike time T , a risk free interest rate r and the function $V(T, S(T))$ from Section 5.1.

1. Use a random number generator to create (approximations of) paths $W(t, \omega_k)$, $k = 1, 2, \dots, N$ of a Wiener process.
2. Apply a numerical method to solve the stochastic differential equation to (approximatively) obtain $x(\tau; t, S(t, \omega_k))$. Set $\tilde{S}_k(T) = x(T; t, S(t, \omega_k))$.
3. Compute the approximation of the expected value via

$$\tilde{E}(V(T, S(T))) = \frac{1}{N} \sum_{k=1}^N V(T, \tilde{S}_k(T)).$$

4. Evaluate the estimate $\tilde{V}(t, S(t)) = \tilde{E}(V(T, S(T))) \exp^{-r(T-t)}$.

Note that for the simple model (5.3), we can utilize the solution formula (5.4) instead of a numerical approximation. To this end, not the entire paths of the Wiener process need to be simulated, only the values $W(T, \omega_k)$ as $\mathcal{N}(0, T)$ -distributed random variables.

Task 5.12 (Monte-Carlo application)

Consider model (5.3)

$$dx(t) = \mu x(t)dt + \sigma x(t)dW_t$$

with $\mu = 0.08$ and $\sigma = 0.2$, initial time $t = 0$ and initial value $S = 80$, and strike price $K = 100$ at time $T = 1$ and suppose the risk free interest rate to be $r = 0.08$. The payoff is given by

$$B = \max\{0, x(T; t, S) - K\}.$$

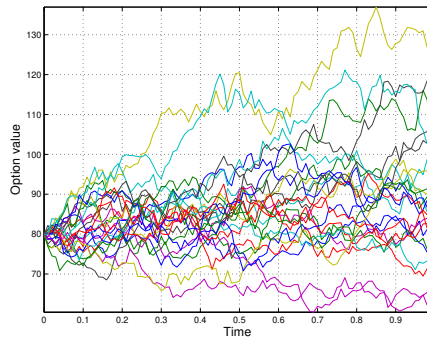
Apply the Monte-Carlo method to approximate $V(t, S(t))$.

Solution to Task 5.12: Applying Algorithm 5.11, we generate 2000 sample paths of the Wiener process in the first step.

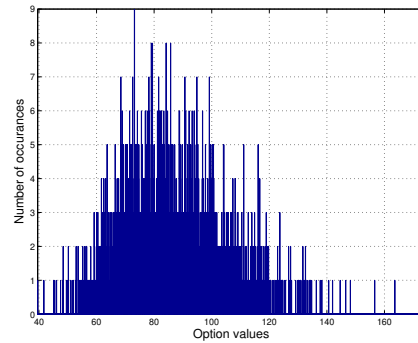
In the second step, we compute the related 2000 solutions of (5.3), cf. Figure 5.1a for a few of these solutions. The large number of samples allows us to approximate the probability density function via a histogram of solutions at strike time T , which is displayed in Figure 5.1b.

Based on these solutions, we can compute the expected value of the underlying stock at

strike time T in the third step. Applying (5.2), we obtain the discounted value of the option displayed in Figure 5.2.



(a) Several solutions paths



(b) Histogram of values $x(T; t, S)$

Figure 5.1.: Numerical results from Task 5.12

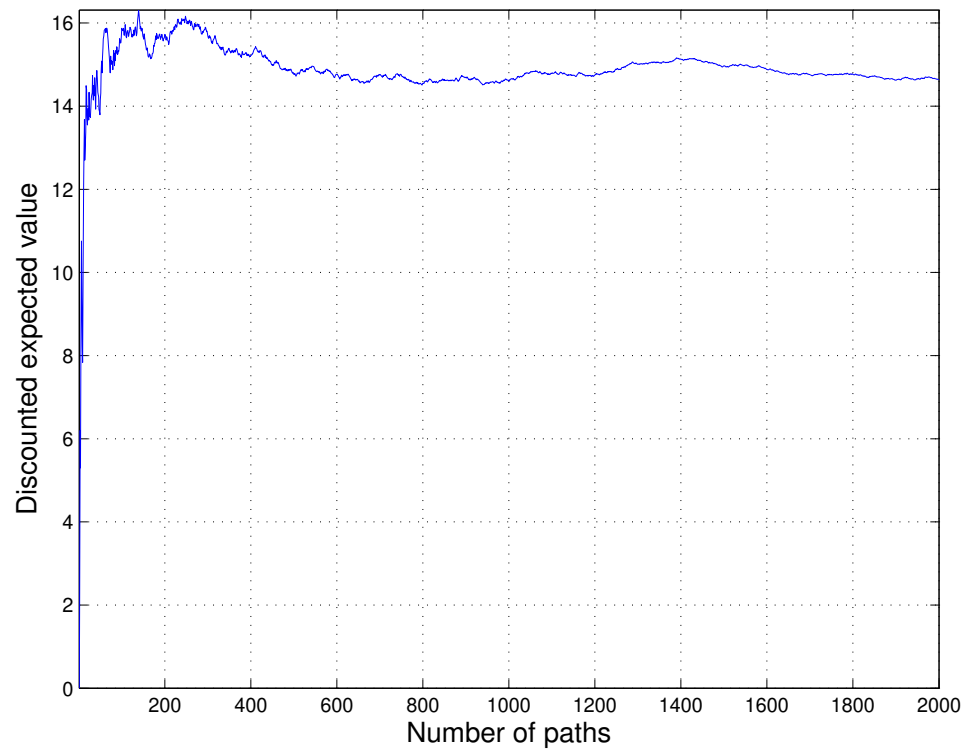


Figure 5.2.: Discounted expected value of the option

The figure illustrates nicely that for large numbers of samples, the solution generated by the Monte–Carlo method converges. Yet, we also observe that quite a large number of samples is required to reduce the fluctuations in the discounted expected value.

To conclude the chapter, we summarize properties of the Monte-Carlo approach to solve stochastic differential equations in Table 5.1.

Table 5.1.: Advantages and disadvantages of Monte-Carlo

Advantage	Disadvantage
✓ Easy to understand	✗ Exhibits slow convergence
✓ Allows to consider complex processes	✗ Computes $S(t)$ for fixed time only
✓ Allows to vary interest rate	

Part II.

Identification

CHAPTER 6

STRUCTURE OF THE IDENTIFICATION PROCESS

Within the Chapter 1, we introduced the notions from stochastic analysis, which we require to study the modeling and identification process. Within the current chapter, we will first discuss the general design sequence of a system identification, which is also called an estimator. Thereafter, we focus on the properties which we are looking for in an estimator. Exemplarily, we will check these properties for a simple electric circuit example. Assessing these estimators will show that there is a clear need for an in deep analysis of properties of estimators.

6.1. Basic design of estimators

As outlined in Chapter 1, every identification consists of the following series of basic steps:

1. Collect information on the system
2. Select a model to represent the system
3. Choose an optimization criterion
4. Fit the model parameters to the measurements accordingly
5. Validate the computed model

Now, we aim to look a bit deeper into each of these steps.

Step 1: Gathering information

In order to identify a process, we first need to build a model of that part of the system, which we are interested in. To this end, we need to gather information about the process. This step

can be done either by observing natural fluctuations, but it is by far more efficient to set up dedicated experiments that actively excite the system via known inputs. While a good example of the first are default fluctuations in demand for a supply chain, the latter can be interpreted as a stress test of a supply chain by uncommon and/or extreme demands. Additionally, the controlled second approach allows for optimization of information gathering goals, such as minimum cost and time, measurement accuracy over a certain bandwidth or other possible aims. Note that the quality of the total identification process may heavily depend on these choices.

Step 2: Selecting the model structure

The model structure is the most variable part of the identification. It not only depends on the problem of identification itself, but may be subject to the further use of the model. For example, an approximation of the elasticity of a wheel via a PDE may give a good dynamical model. Yet, if the model is to be used in a feedback loop, the required computing time to evaluate the model is larger than the sampling time of the loop. Hence, a coarser (or worse) model is necessary for the subsequent task. Keeping this in mind, we distinguish the following:

Parametric vs. nonparametric models

In a parametric model, the system is described by a small number of characteristic quantities. These quantities are called parameters of the model. Regarding our simple electrical circuit example, the expected value is one parameter of the model, the variance the second one. An alternative example is given by the transfer function, e.g. of a filter, which is described by its poles and zeros.

A nonparametric model is given by measurements of a system function at a large number of points. Reviewing the transfer function example, a description via an impulse response at a large number of points is such a characterization.

Note that it is usually simpler to create a nonparametric model than a parametric one because the modeler needs less knowledge about the system itself in the first case. Yet, insight into the problem and concentration of information in a few characteristics is more substantial for parametric models and make the problem simulation faster.

White box vs. black box models

In a white box model, the internal functioning of the system is – at least to some degree – understood. In particular, skills of the experimenter as well as connections between components such as physical laws can be used, whose availability and applicability depend on such an insight. Here, a loudspeaker illustrates the need for extensive understanding of mechanical, electrical and

acoustical phenomena in order to derive an appropriate model.

In contrast to the white box idea of using insight into the system, the black box approach uses a brute force modeling. To this end, a mathematical model is proposed, which allows the description of any observed input and output measurements, but may not even be connected to the real system. Regarding a loudspeaker, a high order transfer function may be used as such a model.

Again, the choice depends on the further aim. While the white box idea provides a better insight into the working principles of the system, the black box model may be sufficient for simulation-/predictions. Note that it is typically a good idea to include as much knowledge as possible during modeling, yet that may not always be easy to accomplish. Analyzing a stable system for example, it is not simple to express this information if polynomial coefficients are used as parameters of the model.

Linear vs. nonlinear models

In almost all cases, real life applications are nonlinear. Unfortunately, theory of nonlinear systems is quite involved and may be difficult to understand for a user unfamiliar with this theory. A nonlinear approach describes the system over its complete operating range and also covers rare and unusual phenomena.

Linear systems, on the other hand, are (almost) completely understood, nice to handle and can be evaluated quickly. Unfortunately, as stated above, real life is typically nonlinear. Therefore, linear systems commonly represent approximations of nonlinear systems within some region – assuming the region can be linearized. Within such a so called operating region, the linear part of the system can be regarded as dominant, i.e. the nonlinear part can be neglected without changing the behavior of the system.

Similar to the other choices, the scope of the problem is relevant to make an appropriate choice. For example, a nonlinear model is needed to describe the distortion of an amplifier, but a linear model is sufficient to represent its transfer characteristics if the operating range is small enough.

Linear-in-parameter vs. nonlinear-in-parameter models

The last choice has to be made between linear and nonlinear influence of parameters of the model. A model is called linear-in-parameter if there exists a linear relation between these parameters and the error that is minimized. Note that linear-in-parameters does not imply a linear model. For example, $e = y - (au^2 + bu + c)$ is linear in a , b and c , but the model is nonlinear. Likewise, $e = y - (a + bj\omega)/(c + dj\omega)u$ is a linear model, but it is nonlinear-in-parameter in c and d .

The impact of this choice can be seen, e.g., for the least square estimator. If the model is linear-in-parameter, then the minimization problem of the least squares can be solved analytically, and does not require an iterative optimization method. Hence, the complexity of a linear-in-parameter

model is much lower.

Step 3: Choose optimization criterion

After choosing a model, it must be matched to the available measurements of the process. To this end, one typically introduces a criterion, which measures the quality of fit, i.e. the distance between the computed and the measured values. Note that since the criterion determines the stochastic properties of the estimator, the choice of this criterion is important for the outcome of the identification process. Regarding our simple resistor example, there are several choices which lead to estimators with different properties.

The cost criterion can be chosen arbitrarily. While its design typically resides on adhoc intuitive insight, there exists a more systematic approach based on stochastic arguments to obtain such a criterion.

Remark 6.1

There exist tests on the cost criterion to check – even before deriving the estimator – if the resulting estimator can be consistent. These are necessary conditions, which are outside the scope of this lecture.

Step 4: Fitting model parameters

In the ensuing step of fitting the parameters, the design work is done and the computations start. Within this step, numerical or symbolic methods are applied to solve the minimization problem arising from the cost criterion in Step 3 subject to the model chosen in Step 2 with respect to the measurement derived in Step 1. Although this step seems to be the essential one, we can already see that the most of the work is the design. This is due to the fact that nowadays, computing power is cheap and there exist a wide area of methods to solve certain problems. The actual art is to design the problem such that it is easily solvable but satisfies the constraints, which bound the model in its further use.

Step 5: Validating obtained model

In the final step, the validity of the obtained model shall be tested. Here, the following question are essential:

- Does the model describe the available data properly?
- Is the model or are parts of it not well designed or flawed?

Note that, as mentioned before, the model with the smallest error is not always the preferred one in practice. Instead, a simpler model may be better suited if it describes the system within user-specified error bounds.

Within the validation process, errors should be separated into different classes such as un-modeled linear dynamics or nonlinearity distortions. Such information shall allow further improvements of the model if necessary. During the validation, the application should be kept in mind, i.e. conditions similar to reality are to be used. Note that extrapolation should be avoided as the errors of extrapolation increase drastically if many measurements are used, which is the typical case for estimator design.

Now that we have seen the general structure of an identification process, we are now interested in properties such an estimator shall offer.

6.2. Properties of estimators

Here, we start of with the claim that a good estimation of a system should exhibit the same characteristics as the system itself, i.e. the same probability density function. Since the probability density function completely defines the properties of a system, such an estimation would do this as well. Unfortunately, without additional conditions it is very hard to show the respective convergence in distribution. But certain properties of the expectation value are sufficient to guarantee mean square convergence, cf. Definition 1.8, which is in turn sufficient for convergence in distribution — the property we like to have.

Within the following, we utilize the assumptions regularly imposed in the literature:

Assumption 6.2 (Noise)

The measurements are disturbed by additive random variables, i.e.

$$y(k) = y_0(k) + X_y(k) \tag{6.1}$$

with the properties that

- each random variable has zero mean and variance σ_y^2 ,
- each random variable is independently and identically distributed (iid),
- each random variable exhibits a symmetric distribution, and
- the random variables are mutually independent.

Hence, our first condition for an estimator is that it reflects an identical expectation value.

Definition 6.3 (Unbiased estimator).

Suppose a probability space (Ω, \mathcal{F}, P) , a measurable space E with σ -algebra \mathcal{E} of E and an estimator (random variable) $\hat{\theta} : \Omega \rightarrow E$ for the parameter $\theta \in E$ to be given. If

$$E(\hat{\theta}) = \theta \quad \forall \theta \in E \quad (6.2)$$

holds true, then we call the estimator $\hat{\theta}$ unbiased. If

$$\lim_{N \rightarrow \infty} E(\hat{\theta}(N)) = \theta \quad \forall \theta \in E \quad (6.3)$$

holds, then we call the estimator $\hat{\theta}$ asymptotically unbiased. Otherwise, it is called biased.

Note that, if the estimator is unbiased, its mean converges towards the mean of the model or model parameters. Yet, since we design the model to represent only a certain part of reality, the model is typically not exact. Hence, the „ideal“ situation is not realistic and we have to think about generalizations. One possibility is to suppose that we evaluate the estimator in a noiseless situation to obtain an approximation. Then, these reference values are compared to results with noise. The final step is to eliminate the influence of the disturbance such that the estimator converges to its reference.

Unfortunately, it is very difficult if not impossible to find the expected value by analytical means. And for some probability density functions, the expected value does not exist. Moreover, we may face the problem that while the expected values are identical, i.e. the estimator is unbiased according to Definition 6.3, the probability density functions are very different and coincide only in the expected value. If such an estimator is used, the outcome of a system may be very different from the real one. To avoid such a problem, we introduce the concept of consistency:

Definition 6.4 (Weak and strong consistency).

Suppose an estimator $\hat{\theta}$ and parameters θ to be given. If $\hat{\theta}$ converges in probability to θ , i.e.

$$\text{p.lim}_{N \rightarrow \infty} \hat{\theta}(N) = \theta, \quad (6.4)$$

then the estimator $\hat{\theta}$ is called weakly consistent.

If $\hat{\theta}$ converges almost surely to θ , i.e.

$$\text{a.s.lim}_{N \rightarrow \infty} \hat{\theta}(N) = \theta, \quad (6.5)$$

then the estimator $\hat{\theta}$ is called strongly consistent.

The advantage of this concept is that we can prove consistency much easier than unbiasedness. If both limits exist, then the limit operator may be interchanged with a continuous function

$$\text{p.lim } f(x) = f(\text{p.lim}(x))$$

and hence the consistency idea exhibits nice calculation properties.

Apart from unbiasedness and consistency, we are also interested in obtaining an estimator, which shows minimal errors only. In particular, we want to minimize the scatter range of the estimator around its limiting value. That gives us the concept of efficiency:

Definition 6.5 (Efficiency).

Suppose an unbiased estimator $\hat{\theta}$ of parameter θ to be given. If for any unbiased estimator $\hat{\theta}_1$ of parameter θ the inequality

$$\text{Cov}(\hat{\theta}, \hat{\theta}) \leq \text{Cov}(\hat{\theta}_1, \hat{\theta}_1) \quad (6.6)$$

holds, then the estimator $\hat{\theta}$ is called efficient.

Since we can rely on a finite number of noisy measurements only, it is clear that there are limits on the accuracy and precision that can be reached by the estimator. The connection between measurements and accuracy is given by the so called Cramer-Rao rule:

Theorem 6.6 (Cramer-Rao rule).

Consider a probability space (Ω, \mathcal{F}, P) and a random variable $X : \Omega \rightarrow E$ defined on that triple, where the set E equipped with measure μ and \mathcal{E} is a σ -algebra of E . Let $f(z, \theta)$ be the probability density function of the measurements $z \in \mathbb{R}^N$. Assume that $f(z, \theta)$ and its first and second derivatives w.r.t. θ exist for all θ and that the boundaries of the domain of $f(z, \theta)$ w.r.t. z are independent of θ . Then, the Cramer-Rao lower bound on the mean square error of any estimator $\hat{G}(\hat{\theta}(z))$ of the function $G(\theta) \in \mathbb{C}^r$ is

$$\text{MSE}(\hat{G}(\hat{\theta}(z))) \geq \left(\frac{\partial G(\theta)}{\partial \theta} + \frac{\partial b_G}{\partial \theta} \right) \text{Fi}(\theta) + \left(\frac{\partial G(\theta)}{\partial \theta} + \frac{\partial b_G}{\partial \theta} \right)^H + b_g b_g^H \quad (6.7)$$

where b_G denotes the expected value bias given by

$$b_G := E(\hat{G}(\hat{\theta}(z))) - G(\theta)$$

and $\text{Fi}(\theta)$ represents the Fisher information matrix of the parameters θ

$$\text{Fi}(\theta) := \text{E} \left(\left(\frac{\partial \ln f(z, \theta)}{\partial \theta} \right)^\top \left(\frac{\partial \ln f(z, \theta)}{\partial \theta} \right) \right) = -\text{E} \left(\frac{\partial^2 \ln f(z, \theta)}{\partial \theta^2} \right).$$

Remark 6.7

We like to stress that the Cramer–Rao rule requires knowledge of the true parameter θ , which may not be at hand. An approximation can still be calculated by replacing θ in (6.7) by its estimated value $\hat{\theta}$. Similarly, the probability density function $f(z, \theta)$ can be approximated using available measurements z only.

The Cramer–Rao rule gives us a very simple way to check efficiency:

Corollary 6.8 (Efficiency).

If a given estimator $\hat{\theta}$ reaches the Cramer–Rao bound (6.7), then it is efficient.

Task 6.9

Consider the electric circuit shown in Figure 6.1 for the resistor model given by Ohm’s law

$$R = u/i. \quad (6.8)$$

Check whether the estimator

$$\hat{R}_{EV}(N) = \frac{\frac{1}{N} \sum_{k=1}^N u(k)}{\frac{1}{N} \sum_{k=1}^N i(k)} \quad (6.9)$$

$$(6.10)$$

is unbiased, consistent and efficient.

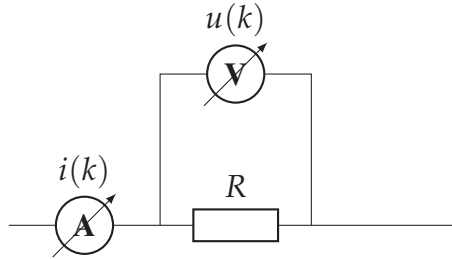


Figure 6.1.: Measurement of a resistor

Solution to Task 6.9:

Unbiasedness: Using the model (6.1) within formula (6.9) we directly see

$$\begin{aligned}
 E(\hat{\theta}_{EV}) &= \lim_{N \rightarrow \infty} \hat{\theta}_{EV}(N) = \lim_{N \rightarrow \infty} \frac{\frac{1}{N} \sum_{k=1}^N u(k)}{\frac{1}{N} \sum_{k=1}^N i(k)} = \lim_{N \rightarrow \infty} \frac{\frac{1}{N} \sum_{k=1}^N u_0 + X_u(k)}{\frac{1}{N} \sum_{k=1}^N i_0 + X_i(k)} \\
 &= \lim_{N \rightarrow \infty} \frac{u_0 + \frac{1}{N} \sum_{k=1}^N X_u(k)}{i_0 + \frac{1}{N} \sum_{k=1}^N X_i(k)} = \frac{u_0 + \lim_{N \rightarrow \infty} \frac{1}{N} \sum_{k=1}^N X_u(k)}{i_0 + \lim_{N \rightarrow \infty} \frac{1}{N} \sum_{k=1}^N X_i(k)}.
 \end{aligned}$$

Now, we can apply the zero mean and iid property of X_u and X_i from our standing Assumption 6.2, that is

$$E(X_u) = \lim_{N \rightarrow \infty} \frac{1}{N} \sum_{k=1}^N X_u(k) = 0, \quad E(X_i) = \lim_{N \rightarrow \infty} \frac{1}{N} \sum_{k=1}^N X_i(k) = 0. \quad (6.11)$$

Hence, we obtain

$$E(\hat{\theta}_{EV}) = \frac{\overbrace{u_0 + \lim_{N \rightarrow \infty} \frac{1}{N} \sum_{k=1}^N X_u(k)}^{=0}}{i_0 + \underbrace{\lim_{N \rightarrow \infty} \frac{1}{N} \sum_{k=1}^N X_i(k)}_{=0}} = \frac{u_0}{i_0} = R_0 \quad (6.12)$$

which shows that the error-in-variable estimator is unbiased.

Consistency: Note that we have already done these computations during the computation of the expected values since we have been using the concept of convergence with probability 1,

which is a stronger concept than convergence in probability. In particular, we have

$$\text{p.lim}_{N \rightarrow \infty} \hat{\theta}_{\text{EV}} = \text{p.lim}_{N \rightarrow \infty} \frac{\frac{1}{N} \sum_{k=1}^N u(k)}{\frac{1}{N} \sum_{k=1}^N i(k)} = \frac{\text{p.lim}_{N \rightarrow \infty} \frac{1}{N} \sum_{k=1}^N u_0 + X_u(k)}{\text{p.lim}_{N \rightarrow \infty} \frac{1}{N} \sum_{k=1}^N i_0 + X_i(k)} = \frac{u_0}{i_0} = R_0.$$

Hence, $\hat{\theta}_{\text{EV}}$ is a weakly consistent estimator.

Efficiency: In order to analyze efficiency, we need to calculate the second moment of it.

Regarding the variance of $\hat{\theta}_{\text{EV}} = R_{\text{EV}}$, we apply Definition 1.5, that is

$$\sigma^2(\hat{\theta}_{\text{EV}}) = \text{E} \left((\hat{\theta}_{\text{EV}} - \text{E}(\hat{\theta}_{\text{EV}}))^2 \right).$$

To compute this value, we reconsider $\hat{\theta}_{\text{EV}}$ and — since we are interested in the second moment only — neglect all second order contributions within such as X_i^2 or $X_u X_i$ in this term, i.e.

$$\begin{aligned} \hat{\theta}_{\text{EV}} &= \frac{u_0 + \frac{1}{N} \sum_{k=1}^N X_u(k)}{i_0 + \frac{1}{N} \sum_{k=1}^N X_i(k)} = \frac{u_0 + \frac{1}{N} \sum_{k=1}^N X_u(k)}{i_0 + \frac{1}{N} \sum_{k=1}^N X_i(k)} \cdot \frac{i_0 - \frac{1}{N} \sum_{k=1}^N X_i(k)}{i_0 - \frac{1}{N} \sum_{k=1}^N X_i(k)} \\ &\stackrel{\text{neglect 2nd order}}{\approx} \frac{u_0 i_0 + \frac{i_0}{N} \sum_{k=1}^N X_u(k) - \frac{u_0}{N} \sum_{k=1}^N X_i(k)}{i_0^2} \\ &= \frac{u_0}{i_0} + \frac{1}{i_0 N} \sum_{k=1}^N X_u(k) - \frac{u_0}{i_0^2 N} \sum_{k=1}^N X_i(k) \\ &= R_0 \left(1 + \frac{1}{N} \sum_{k=1}^N \frac{X_u(k)}{u_0} - \frac{1}{N} \sum_{k=1}^N \frac{X_i(k)}{i_0} \right). \end{aligned}$$

Hence, we obtain

$$\begin{aligned} \sigma^2(\hat{\theta}_{\text{EV}}) &= \text{E} \left((\hat{\theta}_{\text{EV}} - R_0)^2 \right) = \text{E} \left(\left(R_0 \left(\frac{1}{N} \sum_{k=1}^N \frac{X_u(k)}{u_0} - \frac{1}{N} \sum_{k=1}^N \frac{X_i(k)}{i_0} \right) \right)^2 \right) \\ &\stackrel{\text{mutually ind.}}{=} \text{E} \left(\frac{R_0^2}{N^2} \sum_{k=1}^N \frac{X_u(k)^2}{u_0^2} + \frac{R_0^2}{N^2} \sum_{k=1}^N \frac{X_i(k)^2}{i_0^2} \right) \\ &\stackrel{\text{linearity}}{=} \frac{R_0^2}{N^2} \left(\text{E} \left(\sum_{k=1}^N \frac{X_u(k)^2}{u_0^2} \right) + \text{E} \left(\sum_{k=1}^N \frac{X_i(k)^2}{i_0^2} \right) \right) \end{aligned}$$

$$= \frac{R_0^2}{N^2} \left(\frac{\sigma^2(X_u)}{u_0^2} + \frac{\sigma^2(X_i)}{i_0^2} \right) = \frac{R_0^2}{N^2} \left(\frac{\sigma_u^2}{u_0^2} + \frac{\sigma_i^2}{i_0^2} \right)$$

To conclude, we obtain that the error-in-variable estimator $\hat{\theta}_{EV}$ approximate the true value of the parameter and is therefore unbiased. We also found that the error-in-variable estimator converges, yet it is not efficient.

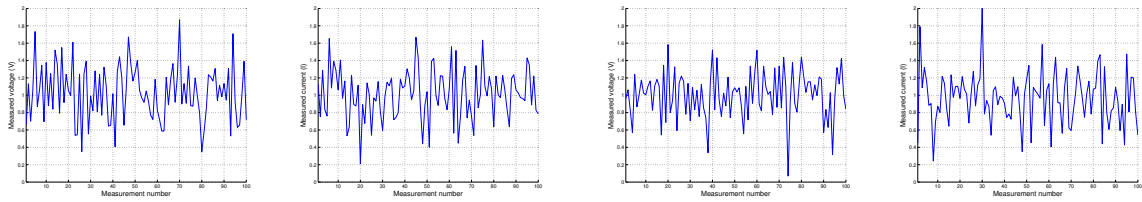
6.3. Practical differences of estimators

In practice, also other estimators such as

$$\hat{R}_{SA}(N) = \frac{1}{N} \sum_{k=1}^N \frac{u(k)}{i(k)} \quad (6.13)$$

$$\hat{R}_{LS}(N) = \operatorname{argmin}_{R \in \mathbb{R}} \sum_{k=1}^N (R \cdot i(k) - u(k))^2 \quad (6.14)$$

may be used. To illustrate the difference, we suppose that two sets of measurements $u(k), i(k)$ with $k = 1, 2, \dots, N$ are given, cf. Figure 6.2.



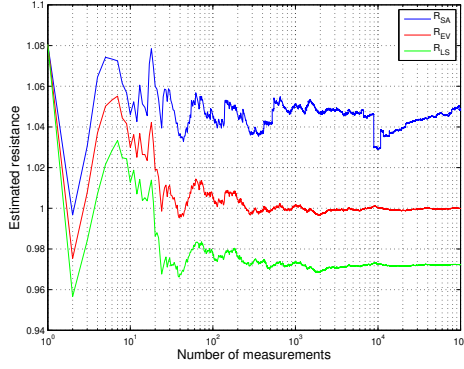
(a) Group A of measurements

(b) Group B of measurements

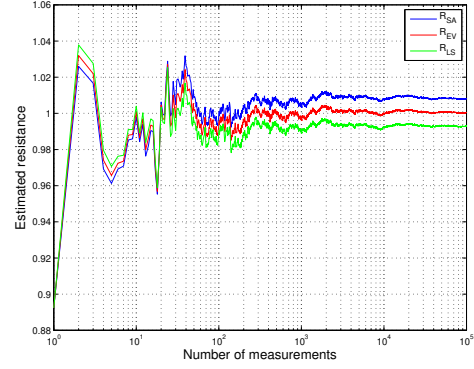
Figure 6.2.: Measurement values for two groups

Utilizing these estimation formulas, we can compute the estimated resistances as displayed in Figure 6.3. From this figure, we can make several observations:

1. All estimators have large variations for small values of N , and – except for \hat{R}_{SA} from group A – show the intuitively expected behavior: for a large number of data points the influence of noise should be eliminated due to the averaging effect.
2. Asymptotic values of the estimators depend on the averaging technique. This shows that the two additional methods converge to a wrong value. Hence, even an infinite amount of measurements does not guarantee that the exact value is found.



(a) Estimated resistance from group A



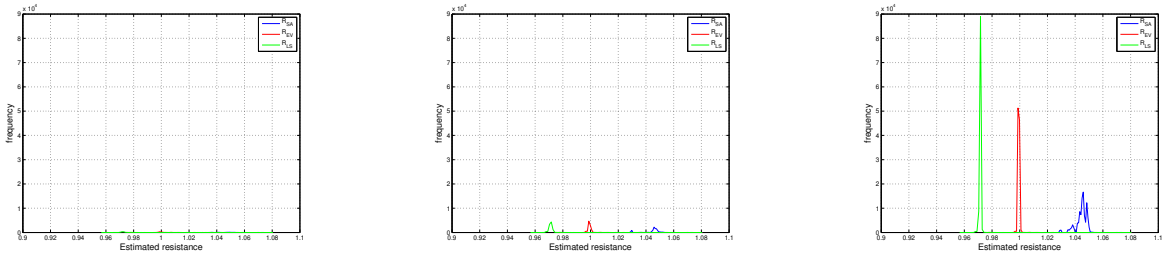
(b) Estimated resistance from group B

Figure 6.3.: Estimated resistances from measurement groups with \hat{R}_{SA} in blue, \hat{R}_{EV} in red and \hat{R}_{LS} in green.

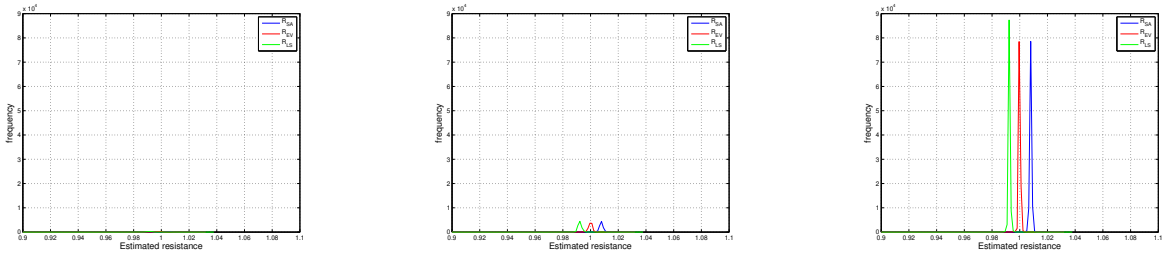
3. \hat{R}_{SA} from group A behaves very strangely. Instead of converging to a fixed value, it jumps irregularly up and down.

These observations indicate that estimators need to be checked regarding their applicability. This allows us to make a sound selection before running expensive experiments.

Continuing a practical approach, we plot approximations of the probability density functions based on the data, cf. Figure 6.4.



(a) Observed probability density functions for group A



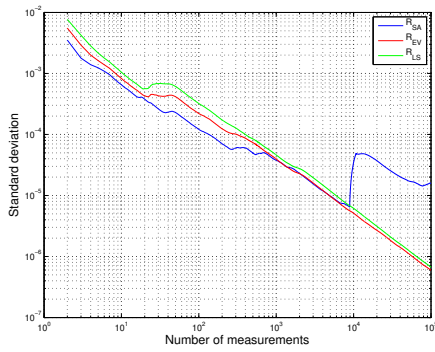
(b) Observed probability density functions for group B

Figure 6.4.: Observed probability density functions for groups. From left to right $N = 1000$, $N = 10000$ and $N = 100000$ with \hat{R}_{SA} in blue, \hat{R}_{EV} in red and \hat{R}_{LS} in green.

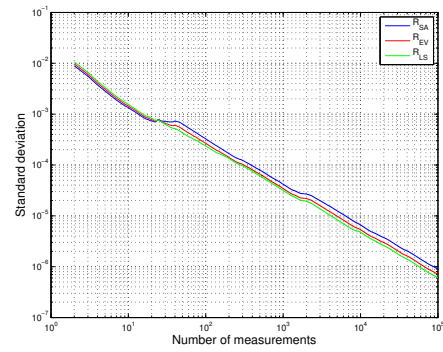
From this figure, we observe the following:

1. For small values of N , the estimates are widely scattered. As the number of processed measurements increases, the probability density function becomes more concentrated.
2. The estimates \hat{R}_{LS} are less scattered than \hat{R}_{EV} and \hat{R}_{SA} , and odd behavior for \hat{R}_{SA} in group A appears again. The distribution of this estimate does not contract to a single value for growing values of N for group A, while it does for group B.
3. The distributions are concentrated around different values.

There seems to be a problem with the measurements of group A, which was observed via \hat{R}_{SA} . In order to quantify the scattering of estimates, in particular of \hat{R}_{SA} , the standard deviation can be calculated, cf. Figure 6.5.



(a) Observed standard deviation for group A



(b) Observed standard deviation for group B

Figure 6.5.: Observed standard deviation for groups. From left to right $N = 1000$, $N = 10000$ and $N = 100000$ with \hat{R}_{SA} in blue, \hat{R}_{EV} in red and \hat{R}_{LS} in green.

Here, we observe that the standard deviation decreases monotonically with N – except for \hat{R}_{SA} of group A. Moreover, the decrease is proportional to $1/\sqrt{N}$, which is the rule of thumb for the uncertainty on an averaged quantity obtained from independent measurements. Additionally, the uncertainty depends on the estimator.

Regarding the strange behavior of \hat{R}_{SA} of group A, we reconsider the measurement data displayed in Figure 6.2 and compute respective histograms, cf. Figure 6.6.

Due to possibly occurring zero values for the current in group A, we obtain a drastic increase in the estimation using the simple approach. This is due to a division by (almost) zero. In group B, such a case does not exist.

Seeing that simply applying estimators is not the best of ideas and may lead to unexpected results, the following Table 6.1 summarizes advantages and disadvantages of theoretically considering properties of estimators before applying these to problems.

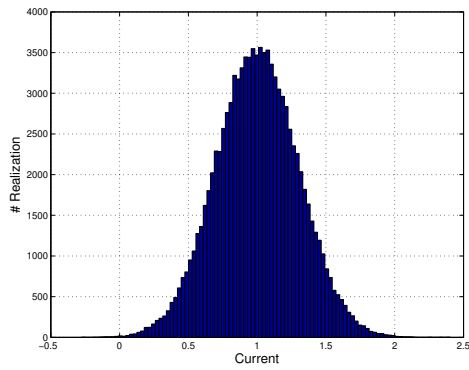
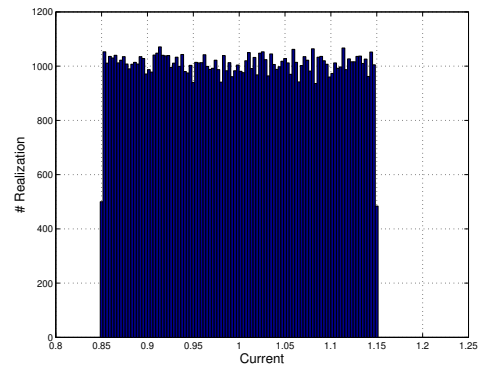
(a) Histogram for $i(\cdot)$ for group A(b) Histogram for $i(\cdot)$ for group BFigure 6.6.: Comparison of histograms for the current $i(\cdot)$

Table 6.1.: Advantages and disadvantages of estimation

Advantages	Disadvantages
✓ Allows to guarantee unbiasedness	✗ May be computable for specific case
✓ Verifies consistency	✗ Not transferable in general
✓ Checks for efficiency	✗ Requires convergence concepts

CHAPTER 7

LEAST SQUARE ESTIMATION

Within the last chapter, we considered the process of identification and properties of estimators, which we would like to have. In this chapter, we pursue a systematic approach to the parameter estimation problem to design an estimator, which satisfies these properties. In particular, we ask what criterion should be used to match the model to the data. To answer this question, we use a statistical approach to select a criterion to measure the quality of the resulting fit. After defining the problem at hand, we discuss two estimators here, the least square and the weighted least square estimator. Note that it is also possible to use other estimator types such as the least absolute values.

7.1. Problem definition

To be rigorous in design, we consider a class of problems that can be defined via inputs and outputs. Note that such systems cover all models we considered in the previous chapters.

Definition 7.1 (Input-output model).

Given a function $g : \mathbb{U} \times \Theta \rightarrow \mathbb{Y}$ we call

$$y_0(k, \theta) = g(u_0(k), \theta) \quad (7.1)$$

input-output system where $k \in \mathbb{N}_0$ represents the measurement index, $y_0(k) \in \mathbb{Y} = \mathbb{R}^{n_y}$ the output, $u(k) \in \mathbb{U} = \mathbb{R}^{n_u}$ the input and $\theta \in \Theta = \mathbb{R}^{n_\theta}$ the vector of true parameters.

The aim is to estimate the parameters θ from noisy observations. To this end, we assume that the output is separated into a deterministic and a probabilistic part $y_0(\cdot)$ and $X_y(\cdot)$.

Assumption 7.2

Given an input output model, noise disturbances only occur within the output observations

$$y(k, X_y) = y_0(k) + X_y(k) \quad (7.2)$$

where $y(k, X_y)$ and $y_0(k)$ represent the modeled and nominal output and $X_y(k)$ denotes the random output variable.

To achieve the described goal, we minimize the errors between simulated and measured values.

Definition 7.3 (Error variable).

Consider an input-output system $g : \mathbb{U} \times \Theta \rightarrow \mathbb{Y}$ to be given and suppose Assumption 7.2 to hold. Then we call

$$e(k, \theta) = z_k - y_0(k, \theta) \quad (7.3)$$

error where z_k denotes the measured (real) output of the system and $y_0(k, \theta)$ denotes the simulated output.

The quality of a fit can then be expressed via a cost criterion. One such criterion is given by the so called nonlinear least squares (NLS), which is derived from the minimization of the sum of squared values:

Definition 7.4 (Least Square estimator).

The least square estimator $\hat{\theta}_{\text{NLS}}(N)$ is given by

$$\hat{\theta}_{\text{NLS}}(N) = \underset{\theta}{\operatorname{argmin}} J_{\text{NLS}}(N, \theta), \quad \text{with } J_{\text{NLS}}(N, \theta) := \frac{1}{2} \sum_{k=1}^N e^2(k, \theta). \quad (7.4)$$

Remark 7.5

Alternatively, one may also use the sum of absolute values

$$\hat{\theta}_{\text{NLA}}(N) = \underset{\theta}{\operatorname{argmin}} J_{\text{NLA}}(N, \theta), \quad \text{with } J_{\text{NLA}}(N, \theta) := \frac{1}{2} \sum_{k=1}^N |e(k, \theta)| \quad (7.5)$$

Since the choice of the cost function we used here is arbitrary, the result only optimal in the sense of this error. Still, least square estimators (7.4) are among the most popular ones, which

is also motivated by numerical aspects: The quadratic nature can be exploited which reveal that the necessary first order conditions for an optimal are also sufficient. We like to mention that the nonlinear least absolute values (7.5) are less sensitive to outliers in the data and may for this reason be interesting in certain applications as well.

As we have seen in Section 6.2, even within the class of least squares different estimators can be designed which lead to results with different properties. In context of an optimal outcome with respect to the properties presented in Chapter 6, it is important to see where the noise enters into the raw data. Thereafter, a cost function should be selected that explicitly accounts for these errors.

7.2. Linear least square

If the model is chosen to be linear-in-parameter θ , equations (7.1) and (7.3) simplify to

Definition 7.6 (Linear-in-parameter input-output system).

Let $g : \mathbb{U} \times \Theta \rightarrow \mathbb{Y}$ define an input-output system. If additionally

$$y_0(\theta) = K(u_0) \theta \quad (7.6)$$

holds with input/output matrix $K(u) \in \mathbb{R}^{N \times n_\theta}$, input vector $u_0 \in \mathbb{R}^N$ and output vector $y_0 \in \mathbb{R}^N$, then it is called *linear-in-parameter input-output system*.

Hence, the error can be rewritten as

$$e(\theta) = z - K(u_0) \theta \quad (7.7)$$

where $z \in \mathbb{R}^N$ represents the vector of measurements. The quality criterion $J_{\text{NLS}}(N, \theta)$ reduces to a linear one given by

$$\begin{aligned} J_{\text{LS}}(N, \theta) &:= \frac{1}{2} e(\theta)^\top e(\theta) = \frac{1}{2} (z - K(u_0) \theta)^\top (z - K(u_0) \theta) \\ &= \frac{1}{2} \sum_{k=1}^N (z_k - K(u_0(k)) \theta)^2. \end{aligned} \quad (7.8)$$

Hence, we obtain the following:

Definition 7.7 (Linear least square estimation problem).

Consider a linear-in-parameter model (7.6) and the linear error function (7.7). Then we call

$$\hat{\theta}_{LS}(N) = \underset{\theta}{\operatorname{argmin}} J_{LS}(N, \theta) \quad (7.9)$$

with $J_{LS}(N, \theta)$ according to (7.8) a linear least square estimator.

Since J_{LS} is quadratic, we can compute the minimizer of this loss function explicitly via

$$\frac{\partial J_{LS}(N, \theta)}{\partial \theta} = 0.$$

This gives us

$$0 = \frac{\partial J_{LS}(N, \theta)}{\partial \theta} = e(\theta)^\top \frac{\partial e(\theta)}{\partial \theta} = e(\theta)^\top (-K(u_0)) = -K(u_0)^\top e(\theta).$$

Hence, we have to solve the equation

$$-K(u_0)^\top (z - K(u_0) \theta) = 0$$

for θ which reveals the solution

$$\hat{\theta}_{LS}(N) = \theta = \left(K(u_0)^\top K(u_0) \right)^{-1} K(u_0)^\top z.$$

Concluding, we have shown the following:

Theorem 7.8 (Solution of linear least square estimator).

The solution to the linear least square estimation problem (7.9)

$$\hat{\theta}_{LS}(N) = \underset{\theta}{\operatorname{argmin}} J_{LS}(N, \theta) \quad \text{with} \quad J_{LS}(N, \theta) = \frac{1}{2} (z - K(u_0) \theta)^\top (z - K(u_0) \theta)$$

is given by

$$\hat{\theta}_{LS}(N) = \left(K(u_0)^\top K(u_0) \right)^{-1} K(u_0)^\top z. \quad (7.10)$$

We like to note that one typically does not compute the least square estimator via formula (7.10),

but instead solves the linear equation

$$\left(K(u_0)^\top K(u_0) \right) \hat{\theta}_{\text{LS}}(N) = K(u_0)^\top z$$

and avoids inverting the matrix $K(u_0)^\top K(u_0)$.

Remark 7.9

Unfortunately, the matrix $K(u_0)^\top K(u_0)$ causes numerical instability since eigenvalues are raised by the power of two. There are, however, ways to compute the solution of the linear least square estimation problem (7.9) by other, more stable algorithms such as the QR decomposition.

In order to generate the matrix K , one has to reformulate the model of the problem (7.6) combined for the available inputs and outputs $u(k)$ and $y(k)$, $k = 1, \dots, N$.

Task 7.10

Consider the model

$$y_0 = \theta,$$

which is independent from the input. Compute the least square estimator for this model.

Solution to Task 7.10: Combining all available outputs $y(k)$, $k = 1, \dots, N$ this reads

$$\begin{aligned} y_0(1) &= \theta \\ &\vdots \\ y_0(N) &= \theta \end{aligned}$$

and reveals the matrix

$$K = \begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix}.$$

Using formula (7.10) we obtain the estimator

$$\hat{\theta}_{\text{LS}}(N) = \left(K^\top K \right)^{-1} K^\top z$$

$$\begin{aligned}
&= \left((1, \dots, 1) \begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix} \right)^{-1} (1, \dots, 1) z \\
&= (N)^{-1} (1, \dots, 1) z = \frac{1}{N} \sum_{k=1}^N z(k).
\end{aligned}$$

The result of Task 7.10 is the average. The result for exemplary model

$$z = \theta \quad \text{with} \quad \theta = 1 + 0.2X_y,$$

where X_y is normally independently distributed with mean 0 and standard deviation 1, i.e. $X_y \in \mathcal{N}(0, 1)$ and $\theta \in \mathcal{N}(1, 0.2)$ is displayed in Figure 7.1 considering 100 measurements.

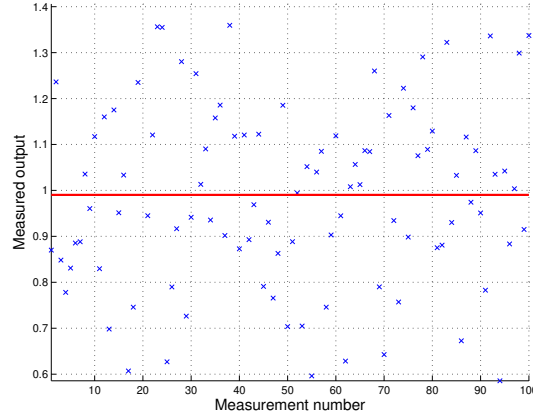


Figure 7.1.: Sample measurements and estimation for Example 7.10

Task 7.11

Consider the model

$$y = u_1\theta_1 + u_2^2\theta_2$$

and compute the respective least square estimator.

Solution to Task 7.11: We can combine inputs and outputs to obtain

$$\begin{aligned} y(1) &= u_1(1)\theta_1 + u_2^2(1)\theta_2 \\ &\vdots \\ y(N) &= u_1(N)\theta_1 + u_2^2(N)\theta_2. \end{aligned}$$

Hence, the linear-in-parameter input-output system reads

$$y_0 = K(u_0) \theta$$

with

$$y_0 = \begin{pmatrix} y(1) \\ \vdots \\ y(N) \end{pmatrix}, \quad u_0 = \begin{pmatrix} u_1(1) \\ u_2(1) \\ \vdots \\ u_1(N) \\ u_2(N) \end{pmatrix}, \quad \text{and} \quad K(u_0) = \begin{pmatrix} u_1(1) & u_2^2(1) \\ \vdots & \vdots \\ u_1(N) & u_2^2(N) \end{pmatrix}.$$

Now, we can apply formula (7.10) to obtain the estimator

$$\begin{aligned} \hat{\theta}_{\text{LS}}(N) &= \left(K(u_0)^\top K(u_0) \right)^{-1} K(u_0)^\top z \\ &= \left(\begin{pmatrix} u_1(1) & u_2^2(1) \\ \vdots & \vdots \\ u_1(N) & u_2^2(N) \end{pmatrix}^\top \begin{pmatrix} u_1(1) & u_2^2(1) \\ \vdots & \vdots \\ u_1(N) & u_2^2(N) \end{pmatrix} \right)^{-1} \begin{pmatrix} u_1(1) & u_2^2(1) \\ \vdots & \vdots \\ u_1(N) & u_2^2(N) \end{pmatrix}^\top z \\ &= \left(\begin{pmatrix} u_1(1) & \dots & u_1(N) \\ u_2^2(1) & \dots & u_2^2(N) \end{pmatrix} \begin{pmatrix} u_1(1) & u_2^2(1) \\ \vdots & \vdots \\ u_1(N) & u_2^2(N) \end{pmatrix} \right)^{-1} \begin{pmatrix} u_1(1) & \dots & u_1(N) \\ u_2^2(1) & \dots & u_2^2(N) \end{pmatrix} z \\ &= \left(\begin{pmatrix} \sum_{k=1}^N u_1^2(k) & \sum_{k=1}^N u_1(k)u_2^2(k) \\ \sum_{k=1}^N u_1(k)u_2^2(k) & \sum_{k=1}^N u_2^4(k) \end{pmatrix} \right)^{-1} \begin{pmatrix} \sum_{k=1}^N u_1(k)z_k \\ \sum_{k=1}^N u_2^2(k)z_k \end{pmatrix}. \end{aligned}$$

The estimator can be computed by solving the two-dimensional linear equation

$$A \cdot \hat{\theta}_{\text{LS}}(N) = b$$

with

$$A = \begin{pmatrix} \sum_{k=1}^N u_1^2(k) & \sum_{k=1}^N u_1(k)u_2^2(k) \\ \sum_{k=1}^N u_1(k)u_2^2(k) & \sum_{k=1}^N u_2^4(k) \end{pmatrix} \quad \text{and} \quad b = \begin{pmatrix} \sum_{k=1}^N u_1(k)z_k \\ \sum_{k=1}^N u_2^2(k)z_k \end{pmatrix}.$$

To illustrate the result, we chose N inputs

$$u_1(k) = 1 + \frac{k}{N-1},$$

$$u_2(k) = 2 + \frac{10k}{N-1},$$

which gives us u_0 and $K(u_0)$. Then, we generated measurements of the form

$$z = K(u_0)\theta$$

with

$$\theta_1 = 1 + 2X_{y,1}$$

$$\theta_2 = 2 + 1X_{y,2}$$

where $X_{y,1}, X_{y,2}$ are normally independently distributed with mean 0 and standard deviation 1, i.e. $\theta_1 \in \mathcal{N}(1, 2)$ and $\theta_2 \in \mathcal{N}(2, 1)$. Considering 100 such measurements, we obtain the result displayed in Figure 7.2.

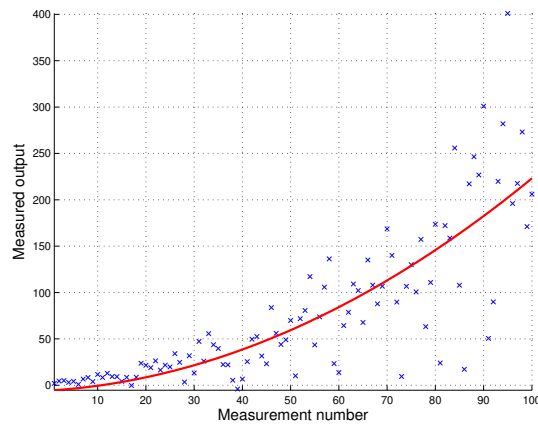


Figure 7.2.: Sample measurements and estimation for Task 7.11

7.3. Properties of the linear least square estimator

Note that we did not formulate any assumptions on the behavior of the noise X_y to compute formula (7.10), but instead calculated it directly from the measurements and the model without bothering about the noise behavior. However, in order to make statements about the properties of the estimator, it is necessary to give some specifications on the noise behavior.

The expected value of the estimator $\hat{\theta}_{LS}$ regarding model outputs, i.e. by considering $z = y(X_y)$, can be computed via

$$\begin{aligned}
 E(\hat{\theta}_{LS}) &\stackrel{(7.10)}{=} E\left(\left(K(u_0)^\top K(u_0)\right)^{-1} K(u_0)^\top y(X_y)\right) \\
 &\stackrel{(7.2)}{=} \left(K(u_0)^\top K(u_0)\right)^{-1} K(u_0)^\top E(y_0 + X_y) \\
 &= \left(K(u_0)^\top K(u_0)\right)^{-1} K(u_0)^\top y_0 + \left(K(u_0)^\top K(u_0)\right)^{-1} K(u_0)^\top E(X_y) \\
 &\stackrel{(7.6)}{=} \left(K(u_0)^\top K(u_0)\right)^{-1} K(u_0)^\top K(u_0) \theta + \left(K(u_0)^\top K(u_0)\right)^{-1} K(u_0)^\top E(X_y) \\
 &= \theta + \left(K(u_0)^\top K(u_0)\right)^{-1} K(u_0)^\top E(X_y).
 \end{aligned}$$

Now, in order for the linear least square estimator to be unbiased, we require $E(X_y) = 0$.

Corollary 7.12 (Unbiasedness of the linear least square estimator).

Consider a linear least square estimator as defined in Definition 7.7. If the probabilistic part of the output satisfies $E(X_y) = 0$, then the least square estimator is unbiased.

The second interesting characteristic is the covariance matrix of the estimator $\hat{\theta}_{LS}$. Here, we see the following:

$$\begin{aligned}
 \text{Cov}(\hat{\theta}_{LS}, \hat{\theta}_{LS}) &= E\left((\hat{\theta}_{LS} - E(\hat{\theta}_{LS}))(\hat{\theta}_{LS} - E(\hat{\theta}_{LS}))^\top\right) \\
 &= E\left(\left(\left(K(u_0)^\top K(u_0)\right)^{-1} K(u_0)^\top E(X_y)\right)\left(\left(K(u_0)^\top K(u_0)\right)^{-1} K(u_0)^\top E(X_y)\right)^\top\right) \\
 &= \left(\left(K(u_0)^\top K(u_0)\right)^{-1} K(u_0)^\top\right) E(X_y X_y^\top) \left(\left(K(u_0)^\top K(u_0)\right)^{-1} K(u_0)^\top\right)^\top \\
 &= \left(\left(K(u_0)^\top K(u_0)\right)^{-1} K(u_0)^\top\right) \text{Cov}(X_y, X_y) \left(\left(K(u_0)^\top K(u_0)\right)^{-1} K(u_0)^\top\right)^\top
 \end{aligned}$$

Similar to Corollary 7.12, we can make the following conclusion regarding the covariance matrix of the estimator $\hat{\theta}_{LS}$.

Corollary 7.13 (Covariance of the linear least square estimator).

Consider a linear least square estimator as defined in Definition 7.7. If the disturbing noise X_y is white and uncorrelated, i.e. $\text{Cov}(X_y, X_y) = \sigma^2(X_y) \text{Id}_{n_\theta}$, then the covariance matrix of the estimator $\hat{\theta}_{LS}$ is given by

$$\text{Cov}(\hat{\theta}_{LS}, \hat{\theta}_{LS}) = \sigma^2(X_y) \left(K(u_0)^\top K(u_0) \right)^{-1} \quad (7.11)$$

Defining $L := \left(K(u_0)^\top K(u_0) \right)^{-1}$, the covariance matrix can be simplified to

$$\text{Cov}(\hat{\theta}_{LS}, \hat{\theta}_{LS}) = L \text{Cov}(X_y, X_y) L^\top. \quad (7.12)$$

Remark 7.14

Here, we like to note that within the least square estimator (7.10)

$$\left(K(u_0)^\top K(u_0) \right) \hat{\theta}_{LS}(N) = K(u_0)^\top z$$

the multiplication $K(u_0)^\top z$ includes an $N \times n_\theta$ and a $n_\theta \times 1$ matrix. To this sum, we can apply the central limit theorem we gives us that the estimator $\hat{\theta}_{LS}$ asymptotically converges to a Gaussian distribution **even** if X_y is not Gaussian distributed, that is

$$\lim_{N \rightarrow \infty} \hat{\theta}_y = \mathcal{N}(\mathbb{E}(\hat{\theta}_{LS}), \text{Cov}(\hat{\theta}_{LS}, \hat{\theta}_{LS})).$$

Task 7.15

Given Assumption 7.2, consider the model from Task 7.10

$$y_0 = \theta$$

and suppose the noise X_y to be white and uncorrelated. Calculate the covariance.

Solution to Task 7.15: Using $K = (1, \dots, 1)^\top$ in (7.11) we obtain

$$\text{Cov}(\hat{\theta}_{LS}, \hat{\theta}_{LS}) = \frac{1}{N} \sigma^2(X_y).$$

Task 7.16

Consider the model from Task 7.11

$$y = u_1\theta_1 + u_2^2\theta_2$$

and again assume Assumption 7.2 to hold and the noise X_y to be white and uncorrelated, i.e. $\text{Cov}(X_y, X_y) = \sigma^2(X_y) \text{Id}_{n_\theta}$. Calculate the covariance matrix.

Solution to Task 7.16: Applying (7.11) we directly have

$$\begin{aligned} \text{Cov}(\hat{\theta}_{\text{LS}}, \hat{\theta}_{\text{LS}}) &= \sigma^2(X_y) \left(K(u_0)^\top K(u_0) \right)^{-1} \\ &\stackrel{\text{Task 7.11}}{=} \sigma^2(X_y) \begin{pmatrix} \sum_{k=1}^N u_1^2(k) & \sum_{k=1}^N u_1(k)u_2^2(k) \\ \sum_{k=1}^N u_1(k)u_2^2(k) & \sum_{k=1}^N u_2^4(k) \end{pmatrix}^{-1}. \end{aligned}$$

Summarizing linear least squares, we obtain the advantages and disadvantages listed in Table 7.1.

Table 7.1.: Advantages and disadvantages of least squares

Advantages	Disadvantages
✓ Direct applicability	✗ Requires „good“ model
✓ Analysis holds for entire method	✗ All inputs treated equally
✓ Unbiasedness and consistency guaranteed under assumptions on stochastic variables	✗ Not efficient

7.4. Weighted least square estimator

So far, we have only been looking at equally weighted measurements in (7.4) (and (7.5)). However, it may be desirable to change this property, e.g. to suppress measurements with high uncertainty and to emphasize those with low uncertainty. To design such a weighting, the covariance matrix can be used.

In practice, it is not always clear which weighting should be used. Yet certain indicators can be used to improve the estimator. For example, if it is known that the model exhibits errors, then utilizing the covariance matrix may not be a good idea. Instead, the user may prefer to put a dedicated weighting in order to keep the model errors small in some specific operation regions.

Definition 7.17 (Weighted Least Square estimator).

Consider a linear-in-parameter model (7.6) and the linear error function (7.7). The weighted least square estimator $\hat{\theta}_{\text{WLS}}(N)$ is given by

$$\hat{\theta}_{\text{WLS}}(N) = \underset{\theta}{\operatorname{argmin}} J_{\text{WLS}}(N, \theta), \quad \text{with } J_{\text{WLS}}(N, \theta) := \frac{1}{2} e(\theta)^\top W e(\theta) \quad (7.13)$$

where $W \in \mathbb{R}^{N \times N}$ is symmetric and positive definite.

Again we can utilize the quadratic nature of J_{WLS} to compute the minimizer explicitly via

$$\frac{\partial J_{\text{WLS}}(N, \theta)}{\partial \theta} = 0.$$

This gives us

$$0 = \frac{\partial J_{\text{WLS}}(N, \theta)}{\partial \theta} = e(\theta)^\top W^\top \frac{\partial e(\theta)}{\partial \theta} = e(\theta)^\top W^\top (-K(u_0)) = -K(u_0)^\top W e(\theta).$$

Solve the equation

$$-K(u_0)^\top W (z - K(u_0) \theta) = 0$$

for θ reveals

$$\hat{\theta}_{\text{WLS}}(N) = \left(K(u_0)^\top W K(u_0) \right)^{-1} K(u_0)^\top W z.$$

Hence, we have shown the following:

Theorem 7.18 (Solution of the weighted linear least square estimator).

The solution to the weighted linear least square estimation problem (7.13) is given by

$$\hat{\theta}_{\text{WLS}}(N) = \left(K(u_0)^\top W K(u_0) \right)^{-1} K(u_0)^\top W z. \quad (7.14)$$

Task 7.19

Consider the model from Task 7.11

$$y = u_1\theta_1 + u_2^2\theta_2.$$

Design a weighting matrix W such that measurements with larger index k are associated with higher weights.

Solution to Task 7.19: One choice could be

$$W = \text{diag}(0, \frac{1}{N-1}, \frac{2}{N-1}, \dots, 1) \in \mathbb{R}^{N \times N}.$$

To illustrate the impact of this choice between the $\hat{\theta}_{\text{LS}}$ and $\hat{\theta}_{\text{WLS}}$, we again reconsider our prior illustration as shown after Task 7.11. For the respective values, we obtain the result display in Figure 7.3. Here, we see that the estimated curve deviates for measurements with small index k . This is to be expected since the respective weights are very small.

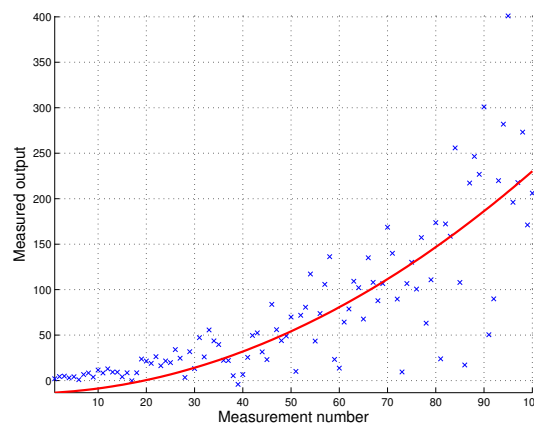


Figure 7.3.: Sample measurements and estimation for Example ??

7.5. Properties of the weighted linear least square estimator

Considering the biasedness, we utilize $z = y(X_y)$ to compute

$$\begin{aligned}
 E(\hat{\theta}_{\text{WLS}}) &\stackrel{(7.14)}{=} E\left(\left(K(u_0)^\top WK(u_0)\right)^{-1} K(u_0)^\top Wy(X_y)\right) \\
 &\stackrel{(7.2)}{=} \left(K(u_0)^\top WK(u_0)\right)^{-1} K(u_0)^\top WE(y_0 + X_y) \\
 &= \left(K(u_0)^\top WK(u_0)\right)^{-1} K(u_0)^\top Wy_0 + \left(K(u_0)^\top WK(u_0)\right)^{-1} K(u_0)^\top WE(X_y) \\
 &\stackrel{(7.6)}{=} \left(K(u_0)^\top WK(u_0)\right)^{-1} K(u_0)^\top WK(u_0)\theta + \left(K(u_0)^\top WK(u_0)\right)^{-1} K(u_0)^\top WE(X_y) \\
 &= \theta + \left(K(u_0)^\top WK(u_0)\right)^{-1} K(u_0)^\top WE(X_y).
 \end{aligned}$$

Now, in order for $\hat{\theta}_{\text{WLS}}$ to be unbiased, we require $E(X_y) = 0$.

Corollary 7.20 (Unbiasedness of the weighted linear least square estimator).

Consider a weighted linear least square estimator as defined in Definition 7.17. If the probabilistic part of the output satisfies $E(X_y) = 0$, then the least square estimator $\hat{\theta}_{\text{WLS}}$ is unbiased.

Similarly, we can compute the covariance matrix of the estimator $\hat{\theta}_{\text{WLS}}$ using the arguments from the unweighted case. Here, we use the abbreviation $K := K(u_0)$.

$$\begin{aligned}
 \text{Cov}(\hat{\theta}_{\text{WLS}}, \hat{\theta}_{\text{WLS}}) &= E\left((\hat{\theta}_{\text{WLS}} - E(\theta))(\hat{\theta}_{\text{WLS}} - E(\theta))^\top\right) \\
 &\stackrel{(7.14)}{=} E\left(\left(\left(K^\top WK\right)^{-1} K^\top WX_y\right)\left(\left(K^\top WK\right)^{-1} K^\top WX_y\right)^\top\right) \\
 &= \left(\left(K^\top WK\right)^{-1} K^\top W\right) E(X_y X_y^\top) \left(\left(K^\top WK\right)^{-1} K^\top W\right)^\top \\
 &= \left(\left(K^\top WK\right)^{-1} K^\top W\right) \text{Cov}(X_y, X_y) \left(\left(K^\top WK\right)^{-1} K^\top W\right)^\top
 \end{aligned}$$

Hence, we can conclude the following about the covariance of $\hat{\theta}_{\text{WLS}}$:

Corollary 7.21 (Covariance of the weighted linear least square estimator).

Consider a weighted linear least square estimator as defined in Definition 7.17. Then the covariance matrix of the estimator $\hat{\theta}_{\text{WLS}}$ is given by

$$\text{Cov}(\hat{\theta}_{\text{WLS}}, \hat{\theta}_{\text{WLS}}) = L \text{Cov}(X_y, X_y) L^\top \quad (7.15)$$

$$\text{where } L := \left(K(u_0)^\top W K(u_0) \right)^{-1} K(u_0)^\top W.$$

This result allows for a very interesting conclusion shown in [1], namely that the covariance matrix can be minimized if the weight is chosen as the inverse of the covariance matrix of the random variable X_y , that is $W = \text{Cov}(X_y, X_y)^{-1}$.

Corollary 7.22 (Minimal covariance of the weighted linear least square estimator).

Consider a weighted linear least square estimator as defined in Definition 7.17. If the weighting matrix is chosen as $W = \text{Cov}(X_y, X_y)^{-1}$, then the covariance matrix of $\hat{\theta}_{WLS}$ is minimal and given by

$$\text{Cov}(\hat{\theta}_{WLS}, \hat{\theta}_{WLS}) = \left(K(u_0)^\top W K(u_0) \right)^{-1}. \quad (7.16)$$

Table 7.2.: Advantages and disadvantages of weighted least squares

Advantages	Disadvantages
✓ Direct applicability	✗ Requires „good“ model
✓ Unbiasedness, consistency and efficiency guaranteed under assumptions on stochastic variables	✗ Requires knowledge on stochastic variable
✓ Analysis holds for entire method	✗ Difficult to adapt

CHAPTER 8

KALMAN FILTERING

Previously, we followed the idea to handle all available data at the same time. This is typically only possible after all measurements have been done, i.e. not at runtime of the process itself. In contrast to that, recursive identification methods aim to iteratively update the estimate utilizing new measurements at hand. Following this approach, an online processing of the results is possible.

Within this chapter, we first introduce the generic concept of recursive identification and its components. Thereafter, we discuss basic properties of the Kalman filter and present the respective algorithm.

8.1. Recursive identification

A straightforward solution to generate such an update procedure is to redo all the calculations after each sample. Such an approach is numerically robust and requires no further insight, yet it may be computationally expensive depending on the number of samples and the complexity of the computation process. For example, it is simple to recompute the mean value, but it is a complex task to solve a nonlinear optimization problem for a dynamical model. Hence, reformulating the problem such that only the newly required calculations are made, recuperating all the previous results, may allow us to generate a more efficient solution method.

Here, we illustrate this approach by considering example of the mean value computation

$$\hat{\theta}(N) = \frac{1}{N} \sum_{k=1}^N z_k.$$

Using this formula, we can recompute the mean value once a new measurement is available via

$$\hat{\theta}(N+1) = \frac{1}{N+1} \sum_{k=1}^{N+1} z_k.$$

To recuperate the previous sum, we can equivalently evaluate

$$\begin{aligned} \hat{\theta}(N+1) &= \frac{1}{N+1} \sum_{k=1}^N z_k + \frac{1}{N+1} z_{N+1} \\ &= \frac{N}{N+1} \hat{\theta}(N) + \frac{1}{N+1} z_{N+1}. \end{aligned}$$

Although this form already meets our requirements of reusing previous computations, it is possible to rearrange it to a more suitable expression:

$$\hat{\theta}(N+1) = \hat{\theta}(N) + \frac{1}{N+1} (z_{N+1} - \hat{\theta}(N))$$

Although this expression is very simple, it is very informative because almost every recursive algorithm can be reduced to a similar form.

For this very special case where the dynamics is linear and the parameter can be measured directly, we obtain the following

Corollary 8.1 (Recursive estimation of mean).

Consider the input–output model

$$y_0 = \theta$$

together with the estimator

$$\hat{\theta}(N) = \underset{\theta}{\operatorname{argmin}} J_{LS}(N, \theta) = \frac{1}{N} \sum_{k=1}^N z_k$$

and suppose Assumption 7.2 to hold. Moreover, suppose measurement data z_k for $k = 1, \dots, N+1$ to be given. Then we have

$$\hat{\theta}(N+1) = \hat{\theta}(N) + \frac{1}{N+1} (z_{N+1} - \hat{\theta}(N)). \quad (8.1)$$

While being very limited in its range of application, the latter result still shows all components we will see for filters. In particular, we observe the following:

- The new estimate $\hat{\theta}(N+1)$ equals the old estimate $\hat{\theta}(N)$ plus a correction term, that is $\frac{1}{N+1} (z_{N+1} - \hat{\theta}(N))$.
- The correction term consists of two terms by itself: a gain factor $\frac{1}{N+1}$ and an error term.
- The gain factor decreases towards zero as more measurements are already accumulated in the previous estimate. This means that in the beginning of the experiment, less importance is given to the old estimate $\hat{\theta}(N)$, and more attention is paid to the new incoming measurements. When N starts to grow, the error term becomes small compared to the old estimate. The algorithm relies more and more on the accumulated information in the old estimate $\hat{\theta}(N)$ and it does not vary it that much for accidental variations of the new measurements. The additional bit of information in the new measurement becomes small compared with the information that is accumulated in the old estimate.
- The second term $z_{N+1} - \hat{\theta}(N)$ is an error term. It incorporates the difference between the predicted value of the next measurement on the basis of the model and the actual measurement z_{k+1} .
- When properly initiated, i.e. $\hat{\theta}(1) = z_1$, this recursive result is exactly equal to the non recursive implementation. However, from a numerical point of view, it is a very robust procedure as calculation errors etc. are compensated in each step.

8.2. Filter problem and assumptions

Broadening the class of problems we just treated, we next consider filter methods. These methods address systems given by dynamics rather than input–output systems. Our aim is to derive the so called Kalman filter. Here, we focus on the discrete time version which is applied to systems given by

$$\begin{aligned} x(k+1) &= Ax(k) + Bu(k) + X_x(k) \\ y(k) &= Cx(k) + X_y(k), \end{aligned} \tag{8.2}$$

where x , u , X_x , y and X_y are vectors and A , B and C are matrices, see also Figure 8.1 for a corresponding block diagram.

Similar to our previous methods, we still consider the error between measurements z and outputs y , but now we aim to identify the state of the system x . In order to classify the Kalman filter problem, we first require a formal distinction of problems regarding information and time dependency.

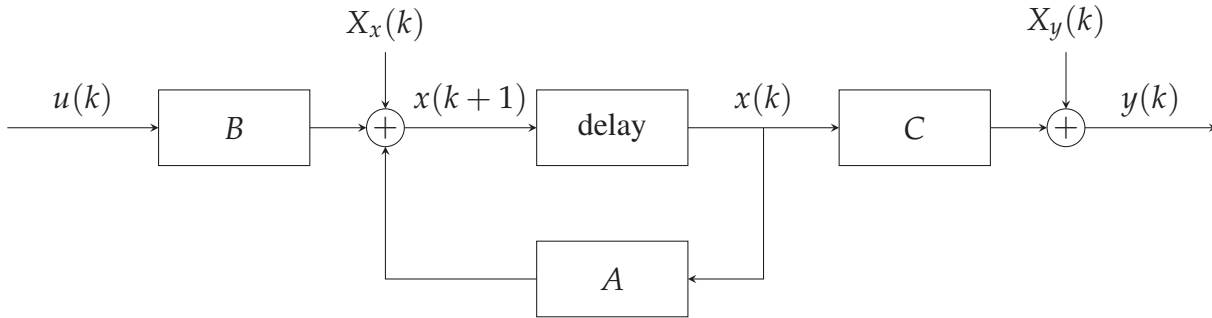


Figure 8.1.: Block diagram of the state space system (8.2)

Definition 8.2 (Filtering).

Consider $x(\cdot)$ to be a state trajectory of a system. Given a specific time instant k , we call the problem of computing

- $x(j)$ with $j < k$ an interpolation problem,
- $x(j)$ with $j = k$ a filtering problem, and
- $x(j)$ with $j > k$ an prediction (or extrapolation) problem.

The Kalman filter itself is more specific as generic filtering. In particular, within the Kalman filter the estimate is computed based on current information and an internal dynamic for the mean value such that new information can be integrated recursively. Our approach to deriving the Kalman filter will involve the following steps:

1. First, we discuss a mathematical description of the model dynamics whose states we want to estimate. Here, we focus on LTI state space models of the form (8.2).
2. Next, we implement equations that describe the propagation of the mean and the covariance of the state with time respectively. These equations form a dynamic system by themselves.

Remark 8.3

Note that the Kalman filter provides an estimator for the state of the dynamical system, i.e.

1. *The mean of the state is the Kalman filter estimate of the state.*
2. *The covariance of the state is the covariance of the Kalman filter state estimate.*

Here, we suppose the following to hold:

Assumption 8.4

Regarding system (8.2) we have that

- the matrices A , B and C are known,
- the matrix B satisfies $B = 0$,
- the random variables X_x and X_y are independent variables,
- the probability density functions f_{X_x} and f_{X_y} are normal distributions,
- the expected values satisfy $E(X_x(k)) = 0$ and $E(X_y(k)) = 0$ and
- the covariance matrices are given by

$$\text{Cov}(X_x(k), X_x(j)) = R_x \delta_{kj} \quad \text{and} \quad \text{Cov}(X_y(k), X_y(j)) = R_y \delta_{kj}.$$

To shorten the notation, we introduce the vector

$$Y(k) := \{y(1), \dots, y(k)\}$$

and denote

$$P(k) := \text{Cov}(x(k) \mid Y(k)) = E \left([x(k) - E(x(k) \mid Y(k))] [x(k) - E(x(k) \mid Y(k))]^\top \right)$$

$$Q(k) := AP(k)A^\top + R_x(k).$$

Given this problem setting, we can now start to derive internal dynamic of the Kalman filter.

8.3. Propagation of mean and covariance

To write down the Kalman filter dynamics, we first need to construct the propagation of the mean value and the covariance regarding past information. Casually speaking, we need to know how these properties evolve regarding past information without new measurements, i.e.

$$\begin{aligned} E(x(k+1) \mid Y(k)) &= E(Ax(k) + X_x(k) \mid Y(k)) \\ &= AE(x(k) \mid Y(k)) + \underbrace{E(X_x(k) \mid Y(k))}_{=0} \\ &= AE(x(k) \mid Y(k)). \end{aligned}$$

Hence, we directly obtain the following:

Theorem 8.5 (Mean propagation).

Given a system (8.2) such that Assumption 8.4 holds. Then we have

$$\mathbb{E}(x(k+1) \mid Y(k)) = A\mathbb{E}(x(k) \mid Y(k)). \quad (8.3)$$

Regarding the covariance dynamics, the update is computationally more involved and reveals

$$\begin{aligned} \text{Cov}(x(k+1) \mid Y(k)) &= \text{Cov}(Ax(k) + X_x(k) \mid Y(k)) \\ &= \mathbb{E} \left(\left(Ax(k) + X_x(k) - \underbrace{\mathbb{E}(Ax(k) + X_x(k) \mid Y(k))}_{=A\mathbb{E}(x(k) \mid Y(k))} \right) \right. \\ &\quad \left. \left(Ax(k) + X_x(k) - \underbrace{\mathbb{E}(Ax(k) + X_x(k) \mid Y(k))}_{=A\mathbb{E}(x(k) \mid Y(k))} \right)^\top \right) \\ &= \mathbb{E} \left((Ax(k) + X_x(k) - A\mathbb{E}(x(k) \mid Y(k))) (Ax(k) + X_x(k) - A\mathbb{E}(x(k) \mid Y(k)))^\top \right) \\ &= \mathbb{E} \left(A(x(k) - \mathbb{E}(x(k) \mid Y(k))) (x(k) - \mathbb{E}(x(k) \mid Y(k)))^\top A^\top \right) \\ &\quad + \underbrace{\mathbb{E}(X_x(k)x(k)^\top \mid Y(k))}_{=0} A^\top + A \underbrace{\mathbb{E}(x(k)X_x(k)^\top \mid Y(k))}_{=0} \\ &\quad - \underbrace{\mathbb{E}(X_x(k)\mathbb{E}(x(k) \mid Y(k))^\top \mid Y(k))}_{=0} A^\top + A \underbrace{\mathbb{E}(\mathbb{E}(x(k) \mid Y(k))X_x(k)^\top \mid Y(k))}_{=0} \\ &\quad + \underbrace{\mathbb{E}(\mathbb{E}(X_x(k))\mathbb{E}(X_x(k))^\top \mid Y(k))}_{=\text{Cov}(X_x, X_x)=R_x} \\ &= A\mathbb{E} \left((x(k) - \mathbb{E}(x(k) \mid Y(k))) (x(k) - \mathbb{E}(x(k) \mid Y(k)))^\top \right) A^\top + R_x \\ &= AP(k)A^\top + R_x \end{aligned}$$

which is exactly our abbreviation $Q(k)$ of the covariance dynamics. Hence, we can conclude:

Theorem 8.6 (Covariance propagation).

Given a system (8.2) such that Assumption 8.4 holds. Then we have

$$\text{Cov}(x(k+1) \mid Y(k)) = AP(k)A^\top + R_x = Q(k). \quad (8.4)$$

Now that we know the estimate of the mean value and the covariance, we can move forward to integrate a new measurement.

To derive an update formula of the estimate of the mean value and the covariance, we need to construct the probability density function of $x(k+1)$. The idea here is to compute an estimate of $x(k+1)$ such that the probability of a respective realization after the measurement of $y(k+1)$ is maximal. This probability density function, in turn, requires an extension of Bayes' rule, which can be derived from the conditional probability density functions

$$\begin{aligned} f(a, b, c) &= f(a | b, c) f(b, c) = f(a | b, c) f(b | c) f(c) \\ f(a, b, c) &= f(a, b | c) f(c). \end{aligned}$$

Combining these two equations, we obtain

$$f(a | b, c) = \frac{f(a, b, c)}{f(b | c) f(c)} = \frac{f(a, b | c) f(c)}{f(b | c) f(c)} = \frac{f(a, b | c)}{f(b | c)}.$$

Substituting $a = x(k+1)$, $b = y(k+1)$ and $c = Y(k)$ reveals

$$\begin{aligned} f(x(k+1) | y(k+1), Y(k)) &= \frac{f(x(k+1), y(k+1) | Y(k))}{f(y(k+1) | Y(k))} \\ &= \frac{f(y(k+1) | x(k+1), Y(k)) f(x(k+1) | Y(k))}{f(y(k+1) | Y(k))} \\ &= \frac{f_{X_y}(y(k+1) - CAx(k)) f(x(k+1) | Y(k))}{f(y(k+1) | Y(k))} \end{aligned} \quad (8.5)$$

where in the second line we used $f(b, c) = f(c, b) = f(c | b) f(b)$ and that for given $Y(k)$ we can substitute $x(k+1) = Ax(k)$ in the third line.

Remark 8.7

Within equation (8.5), the left hand side is the so-called „a posteriori“ probability density function of $x(k+1)$, which includes the knowledge obtained from the measurement $y(k+1)$. On the right hand side, we obtain the „a priori“ probability density function and take the latest measurement $y(k+1)$ into account.

In the following part, we are going to determine $x(k+1)$ such that the probability of realizing $x(k+1)$ after the measurement $y(k+1)$ is maximal. Note that we imposed the limitation that the probability density function of the noise X_x and X_y are normal distributions, cf. Assumption 8.4. Since the covariance matrix $\text{Cov}(x(k+1) | Y(k))$ is given by Lemma 8.6 and R_x , R_y are given by Assumption 8.4, the probability density functions f_{X_x} and f_{X_y} are determined completely. The

denominator of (8.5) is independent of $x(k+1)$ and can therefore be considered as constant when finding the maximum. Hence, we have

$$\begin{aligned}
& \max_{x(k+1)} f(x(k+1) \mid y(k+1), Y(k)) = \\
& = \max_{x(k+1)} e^{-\frac{1}{2}(y(k+1) - CAE(x(k) \mid Y(k)))^\top R_y^{-1} (y(k+1) - CAE(x(k) \mid Y(k)))} \\
& \quad \cdot e^{-\frac{1}{2}(x(k+1) - AE(x(k) \mid Y(k)))^\top Q^{-1}(k+1) (x(k+1) - AE(x(k) \mid Y(k)))} \\
& = \max_{x(k+1)} e^{-\frac{1}{2}(x(k+1) - AE(x(k) \mid Y(k)))^\top (Q^{-1}(k+1) + C^\top R_y^{-1} C) (x(k+1) - AE(x(k) \mid Y(k)))}
\end{aligned}$$

From this equation, we directly obtain

$$\text{Cov}(x(k+1) \mid Y(k+1)) = P(k+1) = Q(k+1)^{-1} + C^\top R_y^{-1} C. \quad (8.6)$$

In order to compute the maximizer of $f(x(k+1) \mid y(k+1), Y(k))$, it is sufficient to minimize the exponent of the above expression. Considering the necessary first order condition, we obtain

$$(Q^{-1}(k+1) + C^\top R_y^{-1} C) (x(k+1) - AE(x(k) \mid Y(k))) = 0$$

In order to obtain stationarity of the evolution, we require $x(k+1) = E(x(k+1) \mid Y(k+1))$. Inserting this into the necessary condition reveals

$$\begin{aligned}
& (Q^{-1}(k+1) + C^\top R_y^{-1} C) E(x(k+1) \mid Y(k+1)) \\
& = Q^{-1}(k+1) AE(x(k) \mid Y(k)) + C^\top R_y^{-1} C AE(x(k) \mid Y(k))
\end{aligned}$$

Now, we can use the matrix inverse lemma

$$P = (Q^{-1} + C^\top R_y^{-1} C)^{-1} = Q - QC^\top (CQC^\top + R_y)^{-1} CQ$$

and the relation

$$(Q + C^\top R_y^{-1} C)^{-1} C^\top R_y^{-1} = QC^\top (CQC^\top + R_y)^{-1}$$

to obtain

$$\begin{aligned}
& E(x(k+1) \mid Y(k+1)) = \\
& = AE(x(k) \mid Y(k)) + \underbrace{Q(k+1)C^\top (CQ(k+1)C^\top + R_y)^{-1}}_{=K(k+1)} (y(k+1) - CAE(x(k) \mid Y(k)))
\end{aligned} \quad (8.7)$$

Remark 8.8

Note that equation (8.7) shows exactly the components highlighted for recursive estimation.

Combined, we obtain the so called Kalman filter algorithm in its most basic form:

Theorem 8.9 (Kalman filter for LTI systems without external input).

Consider a LTI model (8.2) and suppose Assumption 8.4 to hold. Moreover, suppose initial matrices R_x , R_y as well as $X(1)$ to be given and set $P(1) = R_x$. If we abbreviate $X(k) := E(x(k) | Y(k))$, then for $k = 1, \dots$ the equations

$$Q(k+1) = AP(k)A^\top + R_x \quad (8.8)$$

$$K(k+1) = Q(k+1)C^\top \left(CQ(k+1)C^\top + R_y \right)^{-1} \quad (8.9)$$

$$P(k+1) = (Id - K(k+1)C) Q(k+1) \quad (8.10)$$

$$X(k+1) = AX(k) + K(k+1) (y(k+1) - CAX(k)) \quad (8.11)$$

resemble the Kalman filter and provide a recursive estimator satisfying mean and covariance propagation as given in Theorems 8.5 and 8.6.

The algorithm contains several factors, which exhibit a good interpretation regarding the computations made earlier in this chapter. Here, the time component plays an important role.

- The matrix $Q(k+1) = P(k+1 | k)$ represents the a priori covariance matrix of $X(k+1) = E(x(k+1) | Y(k))$ using k measurements only.
- Similarly, the matrix $P(k+1)$ corresponds to the a posteriori covariance matrix of $X(k+1) = E(x(k+1) | Y(k+1))$ using $k+1$ measurements.
- Considering the dynamic of the system, the vector $AX(k)$ reveals the extrapolated state variable based on the model dynamics A and k measurements.
- Projecting on the output, the vector $CAX(k)$ represents the expected output given the extrapolated state of the system.

Remark 8.10

Within the algorithm, the matrices Q , P and K are independent of the measurements. Hence, these can be computed beforehand to lower the computational complexity of the filter. Additionally, the method remains usable when the noise is not normally distributed. In that case, however, the solution found by the filter is no longer an optimal one.

Similar to the case defined by Assumption 8.4, we can consider the more general LTI case with external inputs, i.e. $B \neq 0$. Recall, that the remaining assumptions are still in place, that is

Assumption 8.11

Regarding system (8.2) we have that

- the matrices A , B and C are known,
- the random variables X_x and X_y are independent variables,
- the probability density functions f_{X_x} and f_{X_y} are normal distributions,
- the expected values satisfy $E(X_x(k)) = 0$ and $E(X_y(k)) = 0$ and
- the covariance matrices are given by

$$\text{Cov}(X_x(k), X_x(j)) = R_x \delta_{kj} \quad \text{and} \quad \text{Cov}(X_y(k), X_y(j)) = R_y \delta_{kj}.$$

Given these assumptions, the computations displayed before in this chapter can be modified and the following algorithm can be derived:

Theorem 8.12 (Kalman filter for LTI systems with external input).

Consider a LTI model (8.2) and suppose Assumption 8.11 to hold. Moreover, suppose initial matrices R_x , R_y as well as $X(1)$ to be given and set $P(1) = R_x$. If we abbreviate $X(k) := E(x(k) | Y(k))$, then for $k = 1, \dots$ the equations For $k = 1, \dots$ do

$$Q(k+1) = AP(k)A^\top + R_x \tag{8.12}$$

$$K(k+1) = Q(k+1)C^\top \left(CQ(k+1)C^\top + R_y \right)^{-1} \tag{8.13}$$

$$P(k+1) = (Id - K(k+1)C) Q(k+1) \tag{8.14}$$

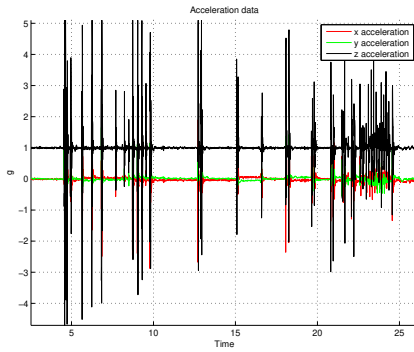
$$X(k+1) = AX(k) + Bu(k) + K(k+1) (y(k+1) - CAX(k) - CBu(k)) \tag{8.15}$$

resemble the Kalman filter and provide a recursive estimator satisfying mean and covariance propagation as given in Theorems 8.5 and 8.6.

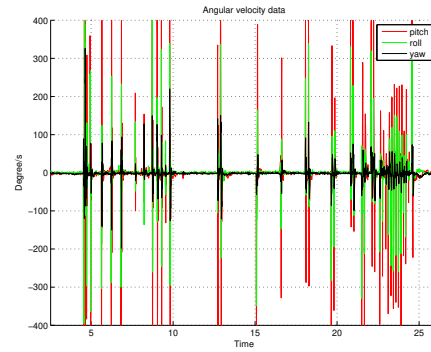
Task 8.13

Consider the data given for a 6DOF inertial measurement unit displayed in Figure 8.2. Given

accelerations along the axes and rotation velocities around the axes, derive the dynamics of a Kalman filter for the sensor fusion problem.



(a) IMU acceleration data in BFC



(b) IMU velocity data in BFC

Figure 8.2.: IMU measurement data from gyros and accelerometers for sudden strikes

Solution to Task 8.13: We define the model dynamics by

$$x(k) = \begin{pmatrix} x_{1,\text{BFC}}(k) \\ \dot{x}_{1,\text{BFC}}(k) \end{pmatrix}, \quad u(k) = \frac{\pi}{180^\circ} \cdot \dot{x}_{1,\text{BFC}}(k)$$

$$A(k) = \begin{pmatrix} 1 & -(t_{k+1} - t_k) \\ 0 & 1 \end{pmatrix}, \quad B(k) = \begin{pmatrix} (t_{k+1} - t_k) \\ 0 \end{pmatrix}, \quad C = \begin{pmatrix} 1 & 0 \end{pmatrix}$$

which gives us the system

$$x(k+1) = \begin{pmatrix} 1 & -(t_{k+1} - t_k) \\ 0 & 1 \end{pmatrix} x(k) + \begin{pmatrix} (t_{k+1} - t_k) \\ 0 \end{pmatrix} u(k) \quad (8.16)$$

$$y(k) = \begin{pmatrix} 1 & 0 \end{pmatrix} x(k) \quad (8.17)$$

To illustrate the results of Task 8.13, we consider the initial value of the Kalman filter

$$x(0) = \begin{pmatrix} \frac{180^\circ}{\pi} \cdot \arctan 2(\ddot{x}_{3,\text{BFC}}, \ddot{x}_{2,\text{BFC}}) \\ 0 \end{pmatrix},$$

and the approximated covariance matrices of the disturbances

$$R_x = \begin{pmatrix} E\left(\frac{\pi}{180^\circ} \cdot 0.0257 \cdot (t_{k+1} - t_k)^2\right) & 0 \\ 0 & 10^{-8} \end{pmatrix}, \quad R_y = \frac{\pi}{180^\circ} \cdot 15,$$

which are based on physical properties of the sensors and a freely chosen bias correction value for $R_{x2,2}$. As a reference, one can also solely use the accelerometer data to evaluate

$$\hat{\theta}_1 = \frac{180^\circ}{\pi} \cdot \arctan 2(\ddot{x}_{3,\text{BFC}}, \ddot{x}_{2,\text{BFC}}). \quad (8.18)$$

Similarly, angular velocity data can be used via integration

$$\hat{\theta}_1(k+1) = \hat{\theta}_{\text{pitch}}(k) + (t_{k+1} - t_k) \dot{x}_{1,\text{BFC}}. \quad (8.19)$$

For both latter approaches, however, we observe that in both cases the angles computed by (8.18), (8.19) diverge, cf. Figure 8.3.

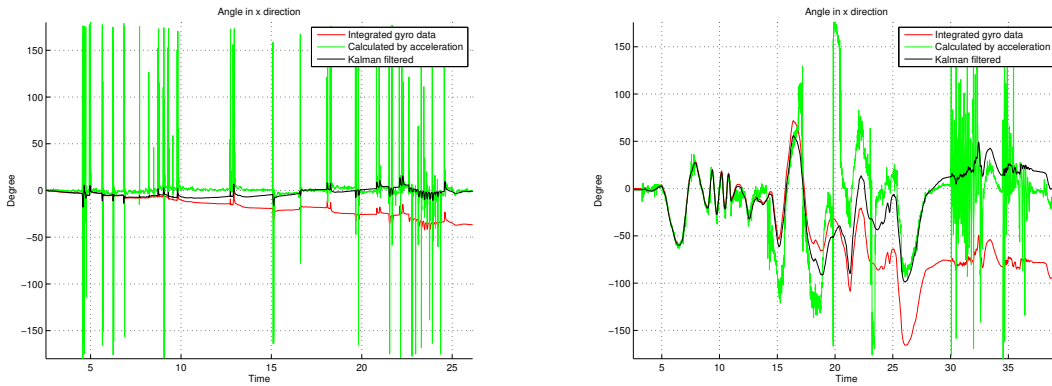


Figure 8.3.: IMU Kalman filter fusion results in comparison to single sensor family results


To summarize, the Kalman filter shows the following advantages and disadvantages:

Table 8.1.: Advantages and disadvantages of Kalman filtering

Advantages	Disadvantages
✓ Applicable to LTI problems	✗ No generic extension to nonlinear case
✓ Resembles recursive estimation	✗ Slow convergence
✓ Allows sensor fusion	✗ Implementation is involved
✓ Predicts/corrects covariance	✗ Requires knowledge on variable

BIBLIOGRAPHY

- [1] AITKEN, A.C.: On Least Squares and Linear Combinations of Observations. In: *Proceedings of the Royal Society of Edinburgh* 55 (1935), pp. 42–48
- [2] AULBACH, B.: *Gewöhnliche Differenzialgleichungen*. Spektrum Akademischer Verlag, 2010
- [3] BORUTZKY, W.: *Bond graph methodology*. Springer, 2009
- [4] HIGHAM, D.J.: *An introduction to financial option valuation: mathematics, stochastics and computation*. Cambridge University Press, 2004
- [5] KHALIL, H.K.: *Nonlinear Systems*. Prentice Hall PTR, 2002
- [6] KLOEDEN, P.E. ; PLATEN, E. ; SCHURZ, H.: *Numerical solution of SDE through computer experiments*. Springer, 2012
- [7] LJUNG, L.: *System Identification: Theory for the User*. Pearson Education, 1998
- [8] MURRAY, J.D.: *Mathematical biology: II: Spatial models and biomedical applications*. Springer, 2003
- [9] PILA, A.W.: *Introduction to Lagrangian dynamics*. Springer, 2019
- [10] SCHOUKENS, J.: *System Identification*. Vrije Universiteit Brussel, 2013
- [11] SCHOUKENS, J. ; PINTELON, R. ; ROLAIN, Y.: *Mastering System Identification in 100 Exercises*. John Wiley & Sons, 2012
- [12] SONTAG, E.D.: *Mathematical Control Theory: Deterministic Finite Dimensional Systems*. Springer, 1998



Jürgen Pannek
Institute for Intermodal Transport and Logistic Systems
Hermann-Blenck-Str. 42
38519 Braunschweig

This script originates from a correspondent lecture *Systemics* held during the summer term 2023 at the Technical University of Braunschweig. To structure the lecture and support my students in their learning process, I prepared these lecture notes.

The aim of the module is to provide participating students with knowledge of terms of system theory and control engineering. Moreover, students shall have knowledge of terms for systems and be enabled to understand principles of system description, modeling and identification. After successfully completing the module, students shall additionally be able to apply the discussed methods and be able to assess results.