AN INTRODUCTION TO DATA BASED MODELING
AN INTRODUCTION TO DATA BASED MODELING

Rik Pintelon

*Vrije Universiteit Brussel*
*dept. ELEC*

Brussels, January 22, 2021
<table>
<thead>
<tr>
<th>Section</th>
<th>Title</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>3.7</td>
<td>Appendix: The Linear Least Squares Estimator is Inefficient</td>
<td>60</td>
</tr>
<tr>
<td>3.8</td>
<td>Appendix: The Weighted Linear Least Squares Estimator is Inefficient</td>
<td>61</td>
</tr>
<tr>
<td></td>
<td>for any Choice $W \neq \lambda C_v^{-1}$</td>
<td></td>
</tr>
<tr>
<td>3.9</td>
<td>Appendix: Proof of the Limit Value $(3.43)$</td>
<td>61</td>
</tr>
<tr>
<td>3.10</td>
<td>Appendix: Weak Consistency of the Total Least Squares Estimator $(3.55)$</td>
<td>61</td>
</tr>
<tr>
<td>3.11</td>
<td>Appendix: Weak Consistency of the Generalized Total Least Squares Estimator $(3.58)$</td>
<td>62</td>
</tr>
<tr>
<td>4</td>
<td>Nonlinear Least Squares</td>
<td>63</td>
</tr>
<tr>
<td>4.1</td>
<td>Problem Statement</td>
<td>63</td>
</tr>
<tr>
<td>4.2</td>
<td>Stochastic Properties</td>
<td>64</td>
</tr>
<tr>
<td></td>
<td>4.2.1 Consistency</td>
<td>65</td>
</tr>
<tr>
<td></td>
<td>4.2.2 Asymptotic Covariance</td>
<td>66</td>
</tr>
<tr>
<td></td>
<td>4.2.3 Asymptotic Efficiency</td>
<td>66</td>
</tr>
<tr>
<td></td>
<td>4.2.4 Asymptotic Normality</td>
<td>66</td>
</tr>
<tr>
<td>4.3</td>
<td>Separable Nonlinear Least Squares</td>
<td>67</td>
</tr>
<tr>
<td></td>
<td>4.3.1 Variable Projection Method</td>
<td>67</td>
</tr>
<tr>
<td></td>
<td>4.3.2 Fitting a Sum of Harmonically Related Sinewaves</td>
<td>68</td>
</tr>
<tr>
<td>4.4</td>
<td>Minimizing the Cost Function</td>
<td>68</td>
</tr>
<tr>
<td></td>
<td>4.4.1 Newton-Raphson</td>
<td>69</td>
</tr>
<tr>
<td></td>
<td>4.4.2 Gauss-Newton</td>
<td>70</td>
</tr>
<tr>
<td></td>
<td>4.4.3 Levenberg-Marquardt</td>
<td>70</td>
</tr>
<tr>
<td></td>
<td>4.4.4 Gradient Descent</td>
<td>71</td>
</tr>
<tr>
<td></td>
<td>4.4.5 Line Search</td>
<td>71</td>
</tr>
<tr>
<td></td>
<td>4.4.6 Stopping Criteria</td>
<td>72</td>
</tr>
<tr>
<td>4.5</td>
<td>Generation of Starting Values</td>
<td>73</td>
</tr>
<tr>
<td></td>
<td>4.5.1 Brute Force Calculations</td>
<td>73</td>
</tr>
<tr>
<td></td>
<td>4.5.2 Convex Approximation</td>
<td>73</td>
</tr>
<tr>
<td>4.6</td>
<td>Appendix: Proof of $(4.9)$ and $(4.10)$</td>
<td>78</td>
</tr>
<tr>
<td>5</td>
<td>Maximum Likelihood Method</td>
<td>79</td>
</tr>
<tr>
<td>5.1</td>
<td>Definition of the Maximum Likelihood Estimator</td>
<td>79</td>
</tr>
<tr>
<td>Section</td>
<td>Title</td>
<td>Page</td>
</tr>
<tr>
<td>---------</td>
<td>--------------------------------------------</td>
<td>------</td>
</tr>
<tr>
<td>5.2</td>
<td>Asymptotic Properties</td>
<td>80</td>
</tr>
<tr>
<td>5.2.1</td>
<td>Consistency</td>
<td>81</td>
</tr>
<tr>
<td>5.2.2</td>
<td>Asymptotic Covariance</td>
<td>83</td>
</tr>
<tr>
<td>5.2.3</td>
<td>Asymptotic Efficiency</td>
<td>83</td>
</tr>
<tr>
<td>5.2.4</td>
<td>Asymptotic Normality</td>
<td>83</td>
</tr>
<tr>
<td>5.2.5</td>
<td>Robustness</td>
<td>84</td>
</tr>
<tr>
<td>5.2.6</td>
<td>Invariance Principle</td>
<td>84</td>
</tr>
<tr>
<td>5.3</td>
<td>Examples</td>
<td>84</td>
</tr>
<tr>
<td>5.3.1</td>
<td>Sample Mean and Sample Variance</td>
<td>84</td>
</tr>
<tr>
<td>5.3.2</td>
<td>Estimation Resistor Value from DC Measurements</td>
<td>87</td>
</tr>
<tr>
<td>5.3.3</td>
<td>Estimation of the Slope of a Straight Line</td>
<td>88</td>
</tr>
<tr>
<td>5.4</td>
<td>Appendix: Proof of (5.48)</td>
<td>93</td>
</tr>
<tr>
<td>6</td>
<td>Bayesian Approach</td>
<td>95</td>
</tr>
<tr>
<td>6.1</td>
<td>Definition of the Bayesian Estimator</td>
<td>95</td>
</tr>
<tr>
<td>6.2</td>
<td>Asymptotic Properties</td>
<td>96</td>
</tr>
<tr>
<td>6.3</td>
<td>Examples</td>
<td>97</td>
</tr>
<tr>
<td>6.4</td>
<td>Gaussian Process Modeling</td>
<td>100</td>
</tr>
<tr>
<td>6.4.1</td>
<td>Estimation Procedure</td>
<td>101</td>
</tr>
<tr>
<td>6.4.2</td>
<td>Gaussian Process Linear Regression</td>
<td>101</td>
</tr>
<tr>
<td>6.4.3</td>
<td>Estimation of a Finite Impulse Response Model</td>
<td>102</td>
</tr>
<tr>
<td>6.5</td>
<td>Appendix: Proof of (6.15)</td>
<td>106</td>
</tr>
<tr>
<td>6.6</td>
<td>Appendix: Proof of (6.17b) and (6.17d)</td>
<td>106</td>
</tr>
<tr>
<td>7</td>
<td>Neural Networks</td>
<td>109</td>
</tr>
<tr>
<td>7.1</td>
<td>Introduction</td>
<td>109</td>
</tr>
<tr>
<td>7.2</td>
<td>Neural Network Models</td>
<td>110</td>
</tr>
<tr>
<td>7.2.1</td>
<td>Components of a Neural Network</td>
<td>110</td>
</tr>
<tr>
<td>7.2.2</td>
<td>Neural Network Topologies</td>
<td>114</td>
</tr>
<tr>
<td>7.2.3</td>
<td>Universal Approximation Property</td>
<td>114</td>
</tr>
<tr>
<td>7.3</td>
<td>Training/Learning a Neural Network</td>
<td>118</td>
</tr>
</tbody>
</table>
Acronyms

DC    Direct Current (used to indicate the zero frequency content of a signal)
EIV   Errors-In-Variables
GP    Gaussian Process
iid   independent and identically distributed
LS    Least Squares
MSE   Mean Squared Error
NLS   Nonlinear Least Squares
NN    Neural Network
pdf   probability density function
QR    QR factorization
RMS   root mean squared
RMSE  root mean squared error
SNR   Signal-to-Noise-Ratio
SVD   Singular Value Decomposition
WLS   Weighted Least Squares
WNLS  Weighted Nonlinear Least Squares
w.p.1 with probability one
w.r.t. with respect to
Operators and Notational Conventions

\[ \tilde{\theta} \]  
estimated value of \( \theta \)

\[ \sigma(A) \]  
singular value of an \( n \times m \) matrix \( A \)

\[ \kappa(A) = \frac{\max_i \sigma_i(A)}{\min_i \sigma_i(A)} \]  
condition number of an \( n \times m \) matrix \( A \)

\[ \|A\|_2 = \max_{1 \leq i \leq \min(n,m)} \sigma_i(A) \]  
2-norm of an \( n \times m \) matrix \( A \)

\[ \mathbb{A} \]  
an outline uppercase font denotes a set, for example, \( \mathbb{N}, \mathbb{Z}, \mathbb{Q}, \mathbb{R}, \) and \( \mathbb{C} \) are, respectively, the natural, the integer, the rational, the real, and the complex numbers

\[ \arg \min_x f(x) \]  
minimizing argument of \( f(x) \)

\[ O(x) \]  
an arbitrary function with the property \( \lim_{x \to 0} |O(x)/x| < \infty \)

\[ o(x) \]  
an arbitrary function with the property \( \lim_{x \to 0} |o(x)/x| = 0 \)

\[ A_{[i,j]} \]  
\( j \)-th column of the matrix \( A \)

\[ A_{[i,:]} \]  
\( i \)-th row of the matrix \( A \)

\[ A_{[n:m,p:q]} \]  
rows \( n \) to \( m \) and columns \( p \) to \( q \) of the matrix \( A \)

\[ X^{[k]} \]  
k-th realization of a random process \( X \)

\[ \sigma_i(A) \]  
i-th singular value of \( A \)
\[ \|X\|_2 = \sqrt{X^T X} \]
\[ \|A\|_F = \sqrt{\text{trace}(A^H A)} \]
\[ \text{diag}(A_1, A_2, \ldots, A_K) \]
\[ \text{std}(x) = \sqrt{\mathbb{E}\{|x - \mathbb{E}(x)|^2\}} \]
\[ \text{var}(x) = \mathbb{E}\{|x - \mathbb{E}(x)|^2\} \]
\[ \text{vec}(A) \]

a. s. lim

l. i. m.

plim

Lim
\[ \mathbb{E}\{\} \]
\[ f^{(n)}() \]
\[ \text{Prob}() \]
\[ \text{Cov}(X, Y) = \mathbb{E}\{(X - \mathbb{E}\{X\})(Y - \mathbb{E}\{Y\})^T\} \]
\[ \text{cov}(x, y) = \mathbb{E}\{(x - \mathbb{E}\{x\})(y - \mathbb{E}\{y\})\} \]
\[ C_X = \text{Cov}(X) = \text{Cov}(X, X) \]
\[ \text{CRLB}(X) \]
\[ \text{Fi}(X) \]
\[ I_m \]
\[ \text{MSE}(X) = \mathbb{E}\{(X - X_0)(X - X_0)^T\} \]
\[ \mu_x = \mathbb{E}\{x\} \]
\[ \sigma_x^2 = \text{var}(x) \]
\[ \sigma_{xy}^2 = \text{covar}(x, y) \]

2-norm of a column vector \( X \)

Frobenius norm of an \( n \times m \) matrix \( A \)

block diagonal matrix with blocks \( A_k, k = 1, 2, \ldots, K \)

standard deviation of \( x \)

variance of \( x \)

column vector formed by stacking the columns of the matrix \( A \) on top of each other

almost sure limit, limit with probability one

limit in mean square

limit in probability

limit in distribution

mathematical expectation

\( n \)-th order derivative of the function \( f() \) w.r.t. its argument

probability

cross-covariance matrix of \( X \) and \( Y \)

covariance of \( x \) and \( y \)

covariance matrix of \( X \)

Cramér-Rao lower bound on \( X \)

Fisher information matrix with respect to the parameters \( X \)

\( m \times m \) identity matrix

mean squared error of \( X \)

expected value of \( x \)

variance of \( x \)

covariance of \( x \) and \( y \)
## Symbols

<table>
<thead>
<tr>
<th>Symbol</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>$e(\theta, z)$</td>
<td>vector of the residuals of the model equation (size $N \times 1$)</td>
</tr>
<tr>
<td>$N$</td>
<td>number of data points</td>
</tr>
<tr>
<td>$n_\theta$</td>
<td>size of the parameter vector $\theta$</td>
</tr>
<tr>
<td>$V(\theta) = \mathbb{E}{V(\theta, z)}$</td>
<td>expected value of the cost function</td>
</tr>
<tr>
<td>$V(\theta, z)$</td>
<td>cost function based on the data vector $z$</td>
</tr>
<tr>
<td>$V'(\theta, z)$</td>
<td>derivative cost function w.r.t. the model parameters $\theta$ (size $1 \times n_\theta$)</td>
</tr>
<tr>
<td>$V''(\theta, z)$</td>
<td>second-order derivative cost function w.r.t. the model parameters $\theta$ (size $n_\theta \times n_\theta$)</td>
</tr>
<tr>
<td>$V_N(\theta) = \mathbb{E}{V_N(\theta, z)}$</td>
<td>expected value of the scaled cost function</td>
</tr>
<tr>
<td>$V_N(\theta, z) = V(\theta, z)/N$</td>
<td>scaled cost function</td>
</tr>
<tr>
<td>$\theta$</td>
<td>vector of the model parameters (size $n_\theta \times 1$)</td>
</tr>
<tr>
<td>$\hat{\theta}(z)$</td>
<td>estimated model parameters; minimizing argument of the cost function $V(\theta, z)$ (size $n_\theta \times 1$)</td>
</tr>
</tbody>
</table>
Chapter 1
Introduction

Abstract: Data-based modeling is the scientific discipline of constructing models from noisy data. It comprises a wide variety of algorithms and models. Although the terminology used depends on the scientific community that is at the origin of a particular approach, the underlying principles are exactly the same. In this chapter we connect the different terminologies and reveal three critical issues in data-based modeling.

Learning Objectives:

- Meaning, goal and terminology of data-based modeling.
- Understanding some critical issues in data-based modeling: (i) the choice of the cost function, (ii) the automatic selection of the model complexity, and (iii) the importance of good starting values for the model parameters to avoid local minima in the cost function.

1.1. data-based Modeling

The ultimate goal of data-based (or data-driven) modeling is to construct a model from noisy data for either prediction, or physical interpretation, control, computer aided design, fault detection, monitoring, classification, denoising of medical images, .... Within the terminology of machine learning, the case we handle in this course is supervised learning, meaning that the data (input-output or output only) is labeled (tagged with properties or characteristics). Starting from one or more noisy data sets (called training sets in machine learning terminology), a model is estimated (trained or learned in machine learning). This model is then used, for example, to predict the response to new inputs, or to detect a faulty operation in new events, or to classify new medical images, ...

In unsupervised learning patterns/features are automatically extracted from noisy data without pre-existing labels. Principal component analysis and clus-
ter analysis are two of the main methods used in unsupervised learning. For example, cluster analysis groups or segments data sets that have not been labeled, categorized or classified before. In classification (= supervised learning), a (non)linear classifier is trained from labeled data, for example, medical images of benign and malignant tumors. Next, the trained (learned) classifier can be used to diagnose new tumor images.

To construct the model from the noisy data, a cost function – which quantifies somehow the distance between the data and the model – is minimized. This step is called estimation in system identification and training or learning in artificial intelligence and machine learning. The choice of the cost function is of key importance because it sets the properties (bias, variability, distribution, ...) of the model obtained. Another important issue of the (automatic) tuning of the complexity of the model. If the model is too complex, then it will also (partially) describe the noise on the data, resulting in a high variability of the estimated (trained or learned) model. This phenomenon is called overmodeling in system identification, overfitting in regression analysis, and overtraining in artificial intelligence and machine learning. Both aspects – the choice of the cost function and the model complexity – are illustrated in the next section.

Two types of models are distinguished: (i) parametric models, which have a fixed number of parameters, and (ii) non-parametric models where the number of parameters increases with the amount of data. An example of a parametric model is a first order differential or state space equation. The impulse response of a discrete-time system is an example of a non-parametric model: as more data samples are collected, more impulse response coefficients are estimated.

1.2. Motivating Examples

In this section we present three simple modeling problems that reveal three critical issues in data-based modeling.

The first motivating example – estimation of the resistor value from DC-measurements – illustrates the importance of choosing an appropriate cost function for constructing a model. It also raises the key question whether we can predict the behavior of the estimated model as more data are collected.

The second motivating example – polynomial curve fitting – illustrates the importance of selecting automatically an appropriate model complexity. A too simple model results in bias errors, while a too complex model leads to a high variability (called poor generalization in neural networks and machine learning) of the predictions on new data.
The third motivating example – estimation of the slope of a straight line – shows the importance of good initial estimates for minimizing a cost function. Without good starting values, there is a high risk to end in a local minimum of the cost function.

### 1.2.1. Estimation of a Resistor Value from DC-Measurements

As an example we take the estimation of the resistor value from DC-measurements taken from Pintelon and Schoukens (2012). The noisy current \( i(k) \) and voltage \( u(k) \) measurements are related to the true values as (see Figure 1.1)

\[
\begin{align*}
  i(k) &= i_0 + n_i(k) \quad (1.1a) \\
  u(k) &= u_0 + n_u(k) \quad (1.1b)
\end{align*}
\]

where \( i_0 = 1 \text{ A} \) and \( u_0 = 1 \text{ V} \). The disturbing measurement noises \( n_i(k) \) and \( n_u(k) \) are assumed to be mutually independent, independent (over the measurement index \( k \)) and identically distributed (iid) zero mean Gaussian random variables

\[
\begin{align*}
  n_i(k) &\sim N(0, \sigma_i^2) \quad (1.2a) \\
  n_u(k) &\sim N(0, \sigma_u^2) \quad (1.2b)
\end{align*}
\]

where \( \sigma_i = 0.5 \text{ A} \) and \( \sigma_u = 0.2 \text{ V} \). \( N = 10^4 \) current and voltage measurements are performed. Figure 1.2 shows the first hundred voltage \( u(k) \) and current \( i(k) \) measurements and the corresponding resistor value obtained via Ohm’s law \( R(k) = u(k)/i(k) \).

From Figure 1.2 it can be seen that the measured resistor value \( R(k) \) has a high variability with now and then large positive and negative peak values.
A basic result from statistics is that averaging decreases the variance by the number of measurements. Applying this principle to the measured resistor values we get

\[ \hat{R}_{SA}(N) = \arg \min_R \sum_{k=1}^{N} \left( \frac{u(k)}{i(k)} - R \right)^2 \]  
\[ = \frac{1}{N} \sum_{k=1}^{N} \frac{u(k)}{i(k)} \]  

The estimate (1.3b) is called the simple approach (SA) and its behavior for increasing values of \( N \) is shown by the red line in Figure 1.3. Surprisingly enough, the averaging procedure 1.3b seems not to decrease the variability of the estimate, preventing convergence for \( N \to \infty \). “Why does the averaging procedure not work?” and “How can we predict this strange behavior?” are important questions that will be answered in Chapter 2.

An alternative solution for estimating the resistor value is to minimize the sum of the squared differences between the measured voltage \( u(k) \) and the voltage \( Ri(k) \) predicted by Ohm’s law

\[ \hat{R}_{LS}(N) = \arg \min_R \sum_{k=1}^{N} \left( u(k) - Ri(k) \right)^2 \]
The behavior of the least squares (LS) estimate (1.4b) for increasing values of $N$ is shown by the green line in Figure 1.3. It can be seen that the variability of the least squares estimate decreases for increasing number of measurements $N$, and the estimate seems to converge to the value $0.8\,\Omega$, which is too small. “Why does the least squares estimator converge?” and “Why does it converge to a value that is too small?” are important questions that will be handled in Chapter 2.

Starting from the probability density function (1.2) of the voltage and current measurements, one can construct the Gaussian maximum likelihood estimator of the resistor value (see Chapter 5)

$$
\hat{R}_{ML}(N) = \arg \min_R \left[ \min_{i_p} \sum_{k=1}^{N} (u(k) - Ri_p)^2 + \sum_{k=1}^{N} (i(k) - i_p)^2 \right]
$$

$$
(1.5a)
$$

$$
= \frac{1}{N} \sum_{k=1}^{N} u(k)
= \frac{1}{N} \sum_{k=1}^{N} i(k)
$$

The behavior of the maximum likelihood (ML) estimate (1.5b) for increasing values of $N$ is shown by the black line in Figure 1.3. It can be seen that the variability of the maximum likelihood estimate decreases for increasing number of measurements $N$, and that the estimate seems to converge to the true value $1\,\Omega$. In a next step, the experiment of gathering hundred current and voltage
measurements is repeated five times, and for each set of hundred current and voltage measurements the corresponding maximum likelihood estimate (1.5b) is calculated for increasing values of \( N \). The results are shown in Figure 1.4. It can be seen that the estimates are different for each of the five experiments, which illustrates that the estimate (1.5b) is a random variable. This should not surprise us because the estimated value depends on the noisy data. Also, for each experiment the estimates seem to converge, and the variability over the experiments seems to decrease for increasing values of \( N \). “Does the maximum likelihood estimator converge for all possible experiments?”, “Does it converge for all experiments to the same (true) value?”, and “Is the observed variability of the estimates the smallest possible?” are important questions that will be discussed in Chapter 2.

From each curve in Figure 1.4, it can be seen that adding a measurement does not necessarily results in an estimate that is closer to the true value of 1 \( \Omega \). At first glance this observation is be very disappointing. However, what we gain by increasing the number of measurements is the decrease in variability resulting in tighter uncertainty bounds.

### 1.2.2. Polynomial Curve Fitting

As curve fitting example we take the polynomial approximation of the arctan function. \( N = 17 \) noisy samples \( y(k), k = 1, 2, \ldots, N \), of the arctan function

![Figure 1.4: Five realizations of the maximum likelihood estimate (1.5b) of resistor value for increasing number \( N \) of measurements.](image)
are generated over the interval $[-2,4]$

$$y(k) = \arctan(u(k)) + n_y(k) \quad \text{with} \quad u(k) = \frac{k - 1}{N - 1} - 2 \quad (1.6)$$

and where $n_y(k) \in N(0, \sigma_y^2)$. These $N$ noisy samples are approximated in least squares sense by a polynomial of order $n_\theta - 1$

$$\hat{\theta} = \arg \min_\theta \sum_{k=1}^{N} \left(y(k) - \sum_{r=1}^{n_\theta} \hat{\theta}_{[r]} u^{(r-1)}(k)\right)^2 \quad (1.7a)$$

$$\hat{y}(k) = \sum_{r=1}^{n_\theta} \hat{\theta}_{[r]} u^{(r-1)}(k) \quad (1.7b)$$

where $\theta_{[r]}$ indicates the $r$-th element of the vector $\theta$. The top row of Figure 1.5 shows the results for $N = 17$, polynomial orders $n_\theta - 1 = 3, 8, 14$, and three
different noise levels $\sigma_y = 0, 0.1, 0.5$. As expected, in the noiseless case, $\sigma_y = 0$, the approximation (bias) error $\hat{y}(k) - \arctan u(k)$ decreases with increasing polynomial order (see the bottom left plot). However, in the noisy case, $\sigma_y \neq 0$, the polynomials of order $n_\theta - 1 = 8$ and 14 have the tendency to follow closely the noisy data points. This is not the case for $n_\theta - 1 = 3$.

Since it is unclear which polynomial order approximates best the arctan function in mean squared sense, the simulation is repeated for $M = 400$ independent noise realizations. Using the $M = 400$ estimates $\hat{y}^{[m]}(k)$, $m = 1, 2, \ldots, M$ and $k = 1, 2, \ldots, N$, (1.7b), the mean squared error can be approximated by its sample estimate

$$\mathbb{E}\{(\hat{y}(k) - \arctan u(k))^2\} \approx \frac{1}{M} \sum_{m=1}^{M} (\hat{y}^{[m]}(k) - \arctan u(k))^2$$  \hspace{1cm} (1.8)

The bottom row of Figure 1.5 shows the square root of (1.8) [for $\sigma_y = 0$ the RMS error is equal to the absolute value of the bias error]. It can be seen that the polynomial order that minimizes the RMS error depends on the noise level: $n_\theta - 1 = 14$ for $\sigma_y = 0$, $n_\theta - 1 = 8$ for $\sigma_y = 0.1$, and $n_\theta - 1 = 3$ for $\sigma_y = 0.5$. This observation can be justified as follows. If the model is too complex, then it also models (partially) the noise. Hence, for each new experiment, the model will change according to the different noise realization, which explains the high variability of the estimated model. Note that the optimal polynomial order is obtained here via minimization of the sample mean squared error, which requires the knowledge of the true function values. How to select automatically the optimal model complexity without this knowledge is an important issue that is addressed in Chapter 8.

1.3. Estimation of the Slope of a Straight Line

We consider here the estimation of the slope of a straight line through the origin from noisy abscissa and noisy ordinate measurements. This is not a special case of the polynomial curve fitting in Section 1.2.2, because it results in a nonlinear minimization problem.

Abscissa $u_0(k)$ and ordinate $y_0(k)$ samples of a straight line through the origin are generated as

$$y_0(k) = \theta_0 u_0(k) \quad \text{with} \quad u_0(k) \in N(0, 0.5^2) \quad \text{for} \quad k = 1, 2, \ldots, N \hspace{1cm} (1.9)$$

and where $\theta_0 = 1$. The slope parameter $\theta_0$ is estimated from $N = 100$ noisy measurements $u(k)$ and $y(k)$ of, respectively, $u_0(k)$ and $y_0(k)$

$$u(k) = u_0(k) + n_u(k) \hspace{1cm} (1.10a)$$
Figure 1.6 shows the cost function (1.12) as a function of the slope parameter $\theta$. It can be seen that the cost function has one local minimum at $\theta = -0.82$ and one global minimum at $\theta = 0.99$ [the true value is one]. Since the minima of (1.12) cannot be found analytically, one of the numerical minimization procedures of Chapter 4 should be applied. All these methods require an initial value for $\theta$ and, if not good enough, one will end up in the local minimum $\theta = -0.82$. Fortunately, by simplifying the problem, we have access to a starting value. Indeed, removing the denominator in each term of (1.12) results in the following linear least squares problem

$$\frac{1}{N} \sum_{k=1}^{N} \frac{(y(k) - \theta u(k))^2}{\sigma_y^2(k) + \theta^2 \sigma_u^2(k)}$$  \hspace{1cm} (1.12)

of which the solution is $\hat{\theta}_{LS} = 0.90$. Although biased, the linear least squares estimate $\hat{\theta}_{LS}$ is close enough to $\hat{\theta}_{ML} = 0.99$ to guarantee the convergence of the numerical minimization procedures to the global minimum of (1.12).
1.4. Outline of the Course

The following topics are handled in this course:

- **Chapter 2**: General tools for analyzing the stochastic properties (bias, variance, convergence, distribution, lower bound on the variance) of estimators when the amount of data increases to infinity.

- **Chapter 3**: Analysis of the stochastic properties (bias, variance) of (weighted) linear least squares estimators for finite amount of data. Numerical stable calculation of the solution. Choice of the weighting. Regularization of the linear least squares solution in case the number of parameters to be estimated is not much smaller than the amount of data. Dealing with noisy regression matrices and handling outliers in measurements.

- **Chapter 4**: Analysis of the stochastic properties (bias, variance, distribution, convergence) of nonlinear least squares estimators when the amount of data tends to infinity. Minimization of the nonlinear least squares costs function: initial value problem, existence of local minima, numerical algorithms (Newton-Raphson, Gauss-Newton, Levenberg-Marquardt, and gradient descent, line search), and numerical stable calculation of the solution.

- **Chapter 5**: Maximum likelihood estimation – principles and basic stochastic properties (bias, variance, distribution, convergence). Properties of the maximum likelihood estimator under non-standard conditions.

- **Chapter 6**: Bayesian estimation – basic principles and properties. Impact of prior knowledge versus data. Gaussian process modeling – relationship with regularized linear least squares. Tuning of the hyper-parameters in Gaussian process regression.


- **Chapter 8**: Automatic tuning of the model complexity. Undermodeling (underfitting), overmodeling (overfitting, overtraining). $L_1$ and $L_2$ regularization, leave-one-out and $p$-fold cross-validation, and penalty terms (AIC, MDL). Use of identification (training) and validation (generalization), and test data sets.
Chapter 2
Tools for Analyzing Estimators

Abstract: Intuitively, we expect from a “good” estimator that it “tends to” the true value as the amount of data increases. How to characterize what is “good”, and what we precisely mean by “tends to” is explained in detail in this chapter.

Learning Objectives:

- Understanding the peculiarities of stochastic convergence.
- Mastering the basic tools for analyzing the asymptotic properties of estimators: the law of large numbers, the central limit theorem, the Cramér-Rao lower bound, and the Taylor series expansion with remainder.
- Understanding the concepts of consistency, asymptotic unbiasedness, asymptotic efficiency, and robustness.
- Getting acquainted with the methodology for proving the asymptotic properties.

2.1. Introduction

First, we describe some basic tools for analyzing the stochastic properties (bias, variance, distribution, convergence, ...) of estimators as the amount of data tends to infinity (Section 2.2). Next, these tools are applied to estimators defined explicitly as a function of the data (Section 2.3), and estimators that are implicitly defined as the minimizing argument of a cost function (Section 2.4).

Ideally, we would like to establish the stochastic properties of estimators for finite data sets. Unfortunately, this is – in general – impossible. What we
hope for is that the asymptotic properties are also valid for a finite amount of data. As a rule of thumb, this is true so long as the number of data points is at least ten times the number of parameters to be estimated. As a good practice, this condition should always be checked beforehand via Monte-Carlo simulations.

2.2. Basic Tools for Analyzing Asymptotic Properties of Estimators

The basic tools presented in this section allows us to answer the following questions: Does the estimate converge for increasing amount of data? To which values does it converge? What is the asymptotic variance and distribution of the estimate? Is it possible to construct an estimator with a smaller variance? Are the asymptotic properties robust with respect to the basic assumptions made to construct the estimator?

2.2.1. Stochastic Convergence

Contrary to the deterministic case, there is no unique definition for the stochastic convergence. In this section we present four different modes of stochastic convergence and discuss their properties and interrelations. All definitions are applied to the scalar random variable \( \hat{\theta} \), but can easily be generalized to the vector case.

A first natural definition is the convergence with probability one. It considers the deterministic limits along all possible realizations of the random variable. If the probability that these limits converge to the same value is equal to one, then the random variable converges with probability one (Stout, 1974; Lukacs, 1975):

\[
\text{Prob}\left( \lim_{N \to \infty} \hat{\theta}(N) = \theta_0 \right) = 1
\]

(2.1a)

where “Prob” denotes probability. Note that \( \theta_0 \) can be a fixed number or a random variable. Note also that (2.1a) does not exclude that the limit might not exist for infinitely many realizations. Alternative notations often used for (2.1a) are

\[
\lim_{N \to \infty} \hat{\theta}(N) = \theta_0 \quad \text{w.p. 1} \quad \text{or} \quad \text{a.s.} \lim_{N \to \infty} \hat{\theta}(N) = \theta_0
\]

(2.1b)

where “w.p.” denotes with probability and where “a.s.” stands for almost sure. Similar to the deterministic limit, the limit with probability one and a continuous function can be interchanged.
A second definition considers the probability that the difference $|\hat{\theta}(N) - \theta_0|$ is arbitrarily small. If, asymptotically, this probability is equal to one, then the random variable converges in probability (Lukacs, 1975):

$$\forall \epsilon > 0 : \lim_{N \to \infty} \text{Prob}(|\hat{\theta}(N) - \theta_0| < \epsilon) = 1$$ (2.2a)

This definition is not equivalent to (2.1a) because – in general – the limit and the probability cannot be interchanged. Alternative notations often used for (2.2a) are:

$$\lim_{N \to \infty} \hat{\theta}(N) = \theta_0 \text{ in prob. \ or } \text{plim}_{N \to \infty} \hat{\theta}(N) = \theta_0$$ (2.2b)

where “p” symbolizes in probability. Similar to the deterministic limit, the limit in probability and a continuous function can be interchanged.

A third definition states that a random variable converges in mean square to some (random) value if the mean squared error of the difference decreases to zero (Jazwinski, 1970):

$$\lim_{N \to \infty} \mathbb{E} \left\{ (\hat{\theta}(N) - \theta_0)^2 \right\} = 0$$ (2.3a)

Note that (2.3a) implies that the variance as well as the bias of the difference decrease to zero, which are natural conditions for convergence. Alternative notations often used for (2.3a) are:

$$\lim_{N \to \infty} \hat{\theta}(N) = \theta_0 \text{ in m.s. \ or \ l.i.m.}_{N \to \infty} \hat{\theta}(N) = \theta_0$$ (2.3b)

where “l.i.m.” stands for limit in mean square. Contrary to the deterministic limit, the limit in mean square and a continuous function cannot be interchanged [2.3a) is a nonlinear operator]. However, the limit in mean square of a linear combination is still the linear combination of the limits in mean square (Jazwinski, 1970).

A fourth definition considers the cumulative distribution functions $F_N(\theta)$ and $F(\theta)$ of the random variables $\hat{\theta}(N)$ and $\theta_0$, respectively. If $F_N(\theta)$ converges to $F(\theta)$ for every $\theta$ at which $F(\theta)$ is continuous, then the random variable converges in distribution or in law:

$$\lim_{N \to \infty} F_N(\theta) = F(\theta)$$ (2.4a)

Alternative notations often used for (2.4a) are:

$$\lim_{N \to \infty} \hat{\theta}(N) = \theta_0 \text{ in distr. \ or \ L}_{N \to \infty} \hat{\theta}(N) = \theta_0$$ (2.4b)

where “L” symbolizes law.
Figure 2.1 visualizes the interrelations between the four stochastic limits by means of a Venn diagram. It can be seen that the almost sure convergence and the mean square convergence imply convergence in probability. All of them imply convergence in law. Note that the convergence in mean square does not necessarily imply convergence with probability one and vice versa. Hence, a.s. lim and l.i.m. are the strongest forms of stochastic convergence, while Lim is the weakest one.

2.2.2. Consistency – Asymptotic Unbiasedness

Using the stochastic convergence criteria (2.1)–(2.3), the consistency of an estimator \( \hat{\theta}(N) \) is defined as

\[
\begin{align*}
\text{strongly consistent : } & \quad \lim_{N \to \infty} \text{a.s. } \hat{\theta}(N) = \theta_0 \\
\text{weakly consistent : } & \quad \lim_{N \to \infty} \text{plim } \hat{\theta}(N) = \theta_0 \\
\text{consistent : } & \quad \lim_{N \to \infty} \text{l.i.m. } \hat{\theta}(N) = \theta_0
\end{align*}
\]

where \( \theta_0 \) is the true value. Conditions (2.5) imply that with high probability \( \hat{\theta}(N) \) is close to \( \theta_0 \). Put differently, the probability mass is concentrated around the true value as the amount of data increases. The choice of the criteria (2.5) depends on the prior knowledge necessary to establish the convergence [see Section 2.2.3 on page 15]. Criterion (2.5c) has the disadvantage that it cannot be interchanged with a continuous function.

An estimator \( \hat{\theta}(N) \) is called asymptotically unbiased if

\[
\lim_{N \to \infty} \mathbb{E}\{\hat{\theta}(N)\} = \theta_0
\]

where \( \theta_0 \) is the true value. It is unbiased if \( \mathbb{E}\{\hat{\theta}(N)\} = \theta_0 \) for \( N \geq n_0 \).
Figure 2.2: Probability density function $f_{\hat{\theta}}(\hat{\theta})$ of a consistent (right) and asymptotically unbiased (left) estimators. [Figure adapted from Pintelon and Schoukens (2012)]

Note that (2.6) does not imply consistency (2.5) and vice versa. The difference between both properties is visualized in Figure 2.2. For a consistent estimator (right plot), the probability mass is concentrated around the true value $\theta_0$, which is not necessarily the case for an asymptotically unbiased estimator (left plot). On the other hand, consistency does not guarantee that the variance and the expected value of the estimator exist. In order to exist, the skirts of the probability density function should decrease sufficiently fast to zero. From this discussion it is clear why consistency is mostly the desired property. However, in some applications (asymptotic) unbiasedness is required.

### 2.2.3. Law of Large Numbers

Consider the mean value $S(N)$ of $N$ random variables $x(k)$, $k = 1, 2, \ldots, N$,

$$S(N) = \frac{1}{N} \sum_{k=1}^{N} x(k)$$  \hspace{1cm} (2.7)

The law of large numbers states under which conditions the mean (2.7) converges in stochastic sense to its expected value. According to the stochastic convergence criterion used one distinguishes

- **strong law of large numbers**: $\lim_{N \to \infty} \mathbb{E}\{S(N) - E\{S(N)\}\} = 0$  \hspace{1cm} (2.8a)
- **weak law of large numbers**: $\lim_{N \to \infty} \mathbb{E}\{S(N) - E\{S(N)\}\} = 0$  \hspace{1cm} (2.8b)
- **law of large numbers**: $\lim_{N \to \infty} \mathbb{E}\{S(N) - E\{S(N)\}\} = 0$  \hspace{1cm} (2.8c)

For example, for independent and identically distributed (iid) random variables $x(k)$, the strong law of large numbers (2.8a) is valid [Theorem 4.3.3 of Lukacs, 1975].
1975]. Note that the variance of \( x(k) \) should not exist. If not, then the mean value (2.7) still converges with probability one [strong law of large numbers (2.8a)], but not in mean square sense [law of large numbers (2.8c)].

For independent random variables \( x(k) \) with finite variances

\[
\text{var}(x(k)) \leq C_1 < \infty
\]  

(2.9)

(2.8a) and (2.8b) are equivalent [Theorem 2.13.2 of Stout, 1974]. Note that condition (2.9) ensure that, asymptotically, infinity many random variables contribute significantly to the mean value. If (2.9) is not be fulfilled, then (2.7) is be dominated by the random variables with asymptotically infinite variance so that no useful average effect occurs.

When measuring the response of dynamical systems, the disturbing noise produced by the system itself – called process noise – has mostly a non-white power spectrum (spectral density). It implies that consecutive disturbing noise samples are correlated. Fortunately, if the correlation decreases sufficiently fast to zero as the time-difference increases (mixing condition of order two), then the law of large numbers (2.8c) is valid [see Appendix 16.G of Pintelon and Schoukens, 2012].

### 2.2.4. Asymptotic Covariance

A condition for an estimator \( \hat{\theta}(N) \) to convergence in stochastic sense to \( \theta_0 \) is that somehow the disturbing noise is averaged. Via a first order Taylor series expansion of \( \hat{\theta}(N) \) around the limit value \( \theta_0 \), an explicit expression is obtained for \( \hat{\theta}(N) - \theta_0 \) as a function of the disturbing noise [see Sections 2.3 and 2.4]. The covariance of this expression can be calculated, and is called – with some abuse of terminology – the asymptotic covariance of \( \hat{\theta}(N) \).

Note that, in general, the covariance of \( \hat{\theta} \) does not exist, even for simple examples. However, the asymptotic covariance matrix combined with the asymptotic normality [see Section 2.2.5] allows us to construct uncertainty bounds on \( \hat{\theta} \) with a given confidence level. This is illustrated in Section 2.3 on the estimation of a resistor value form DC current and voltage measurements.

### 2.2.5. Central Limit Theorem – Asymptotic Normality

Classical central limit theorems pertain to the asymptotic distribution function of the mean value (2.7). They state roughly that \( S(N) \) is asymptotically
normally distributed

\[
\lim_{N \to \infty} \frac{S(N) - \mathbb{E}\{S(N)\}}{\sqrt{\text{var}(S(N))}} \in N(0, 1)
\] (2.10)

if each \(x(k)\) in (2.7) has high probability to be of the same order of magnitude and if the span of dependence of \(x(k)\) is limited. Some interesting versions of the central limit theorem are listed next. More version can be found in Feller (1968), Chow and Teicher (1988) and Billingsley (1995).

1. If \(x(k)\) is independent and identically distributed with finite mean \(\mathbb{E}\{x(k)\} < \infty\) and finite nonzero variance \(0 < \text{var}(x(k)) < \infty\), then (2.10) is valid [Theorem 27.1 of Billingsley, 1995].

2. \(x(k)\) is independent with finite means \(\mathbb{E}\{x(k)\} \leq c_1 < \infty\) and finite variances (2.9). If in addition there exist an \(\epsilon > 0\) such that \(x(k)\) has uniformly bounded \(2 + \epsilon\) moments \(\mathbb{E}\{|x(k)|^{2+\epsilon}\}\) and if

\[
\lim_{N \to \infty} \frac{N}{\sqrt{\text{var}(NS(N))}}^{2+\epsilon} = 0
\] (2.11)

then (2.10) is valid [Theorem 27.3 of Billingsley, 1995].

3. If the \(x(k)\) are correlated then all higher order correlations should decrease sufficiently fast to zero (mixing condition of order infinity). If in addition

\[
\lim_{N \to \infty} \frac{N}{\text{var}(NS(N))} = O(N^0)
\] (2.12)

then (2.10) applies [see Theorem 4.4.1 of Brillinger, 1981 for the stationary case and Appendix 16.H of Pintelon and Schoukens, 2012 for the nonstationary case].

Conditions (2.11) and (2.12) are necessary to avoid dominance of a few random variables in the mean value \(S(N)\). If not, the distribution of \(S(N)\) would be determined by the distribution functions of those few dominating random variables and would – in general – not be normal. Extensions of these theorems to the complex and to the (complex) multivariate case is straightforward [see Brillinger, 1981].

The central limit theorem should be interpreted and applied with some care:
1. Consider the mean (2.7), where the \( x(k) \) are independent and uniformly distributed with zero mean value and variance \( \frac{25}{3} \). Figure 2.3 compares the true probability density function (pdf) of \( S(N) \) to the Gaussian pdf predicted by the central limit theorem. Although the mean \( S(N) \) cannot take values outside the interval \([-5, 5]\), the central limit theorem predicts that this will happen with some (small) probability. Similarly, saying that the weight of newborn babies is normally distributed does not imply that there is a small risk of getting babies with a negative weight. We conclude that the central limit theorem describes very well the behavior of the distribution function around its mean value, but not at its tails.

2. An infinite sum of asymptotically normally distributed random variables is not necessarily normally distributed. The following example illustrates this. Consider the discrete Fourier transform (DFT) \( X(k) \) of \( N \) uniformly distributed samples \( x(t), t = 0, 1, \ldots, N - 1 \).

\[
X(k) = \sum_{t=0}^{N-1} x(t)e^{-2\pi j \frac{kt}{N}} \tag{2.13}
\]

According to the central limit theorem, \( X(k) \) is asymptotically normally distributed (Brillinger, 1981). Taking the inverse discrete Fourier transform (IDFT) of \( X(k) \),

\[
x(t) = \frac{1}{N} \sum_{k=0}^{N-1} X(k)e^{2\pi j \frac{kt}{N}} \tag{2.14}
\]

Figure 2.3: Comparison of the true pdf (solid lines) of \( S(N) \) (2.7) and the Gaussian pdf (dashed lines) predicted by the central limit theorem (2.10), for zero mean, independent distributed random variables \( x(k) \), with \( \text{var}(x(k)) = 25/3 \): (a) \( N = 1 \), (b) \( N = 2 \), and (c) \( N = 3 \). [Figure taken from Pintelon and Schoukens (2012)]
we recover the uniformly distributed samples \( x(t) \). The third version of the central limit theorem cannot be applied to the \( X(k) \) in (2.14) because, although uncorrelated over \( k \), the higher order correlations do not decrease sufficiently fast to zero (not mixing of order infinity). Hence, using the DFT-IDFT pair of transforms, we can generate infinitely many asymptotically normally distributed random variables \( X(k) \) for which the central limit theorem does not hold.

### 2.2.6. Cramér-Rao Lower Bound – Asymptotic Efficiency

Consider the estimation of the parameter vector \( \theta \in \mathbb{R}^n \) using noisy measurements \( z \in \mathbb{R}^N \) with probability density function \( f_z(z) \). From \( f_z(z) \) and the model, we can derive the probability density of the data \( z \), given the true model parameters \( \theta_0 \): \( f_{z|\theta_0}(z|\theta_0) \). Assume now that (i) the first (and second) order derivative of \( f_{z|\theta_0}(z|\theta_0) \) w.r.t. \( \theta_0 \) exists for all \( \theta_0 \)-values, and (ii) the boundaries of the domain of \( f_{z|\theta_0}(z|\theta_0) \) w.r.t. \( z \) are \( \theta_0 \)-independent. Within the class of unbiased estimators, the covariance of the estimate \( \hat{\theta}(z) \) is then bounded below by the Cramér-Rao lower bound

\[
\text{Cov}(\hat{\theta}(z)) \geq \text{Fi}^{-1}(\theta_0)
\]

(2.15)

where \( \text{Fi}(\theta_0) \) is the Fisher information matrix defined as

\[
\text{Fi}(\theta_0) = \mathbb{E}\left\{ \left( \frac{\partial \log f_{z|\theta_0}(z|\theta_0)}{\partial \theta_0} \right)^T \left( \frac{\partial \log f_{z|\theta_0}(z|\theta_0)}{\partial \theta_0} \right) \right\} \quad (2.16a)
\]

\[
= -\mathbb{E}\left\{ \frac{\partial^2 \log f_{z|\theta_0}(z|\theta_0)}{\partial \theta_0^2} \right\} \quad (2.16b)
\]

(proof: see Appendix 2.6 on page 33).

Discussion of the lower bound (2.15):

1. The lower bound (2.15) implies that, within the class of unbiased estimators, the covariance of \( \hat{\theta}(z) \) can never be smaller than the inverse of the Fisher information matrix.

2. The calculation of the Cramér-Rao lower bound requires the knowledge of the true parameter value \( \theta_0 \) and the noiseless data \( z_0 \), which are not available except in simulations. In practice, an approximation of (2.15) is calculated by replacing \( \theta_0 \) and \( z_0 \) by, respectively, \( \hat{\theta}(z) \) and \( z \).
3. The lower bound (2.15) may be too conservative. Better (larger) bounds exist when (2.15) is not attainable but they are often (extremely) difficult to calculate. An overview of tighter bounds can be found in Abel (1993).

4. While (2.15) is valid for finite $N$ within the class of unbiased estimators, it is asymptotically ($N \to \infty$) valid for asymptotically ($N \to \infty$) unbiased estimators and consistent estimators. In the case of consistent estimators, the covariance of the limit random variable the estimator is converging to, is used in the lower bound.

5. For uniformly distributed measurements $z$, the borders of the domain of $f_{z|\theta_0}(z|\theta_0)$ are not $\theta_0$-independent and, hence, the Cramér-Rao lower bound (2.15) does not exist.

An (asymptotically) unbiased estimator is called (asymptotically) efficient if the equality holds (asymptotically) in (2.15). A consistent estimator is called asymptotically efficient if the covariance of the limit random variable it is converging to, reaches the lower bound (2.15).

2.2.7. Robustness

Estimators are constructed under some assumptions on the data. For example, the probability density function of the disturbing noise, the independence of measured samples, the prior knowledge about the parameters and/or the disturbing noise, ... In practice, these assumptions are not perfectly met. Hence, it is important to verify how sensitive the asymptotic properties of an estimator are w.r.t. deviations from the basic assumptions made.

The robustness analysis basically boils down to verifying whether the tools used for proving the asymptotic properties (law of large numbers, central limit theorem, Taylor series expansion, asymptotic covariance, Cramér-Rao lower bound) are still valid under relaxed conditions. This is illustrated in Section 2.3 on the estimation of the resistor value from DC current and voltage measurements.

2.3. Analysis of Estimators Defined Explicitly as a Function of the Noisy Data

In this section we analyze the asymptotic ($N \to \infty$) properties of the simple approach (1.3), the least squares estimator (1.4) and the maximum likelihood
solution (1.5) for estimating the resistor value from noisy DC current and voltage measurements. Each asymptotic property will be handled separately.

### 2.3.1. Consistency

Since the current measurements are independent and identically distributed and similarly for the voltage measurements, we can apply the strong law of large numbers (2.8a) to each of the sums in (1.3b), (1.4b) and (1.5b). Hence, to calculate the limit value of each of the estimates, we have to calculate the expected value of each sum.

We find for the simple approach (1.3b)

\[
\hat{R}_{SA}(N) \xrightarrow{w.p.1} \frac{1}{N} \sum_{k=1}^{N} \mathbb{E}\left\{\frac{u(k)}{i(k)}\right\} = \frac{1}{N} \sum_{k=1}^{N} \mathbb{E}\{u(k)\}\mathbb{E}\left\{\frac{1}{i(k)}\right\} = \infty \quad (2.17)
\]

where the first equality uses \(\mathbb{E}\{xy\} = \mathbb{E}\{x\}\mathbb{E}\{y\}\) for independent random variables \(x\) and \(y\), and the second equality \(\mathbb{E}\{1/x\} = \infty\) for a Gaussian random variable \(x\). Result (2.17) explains why the simple approach in Figure 1.3 seems not to converge.

For the least squares method (1.4b) we get

\[
\hat{R}_{LS}(N) \xrightarrow{w.p.1} \frac{1}{N} \sum_{k=1}^{N} \mathbb{E}\left\{u(k)i(k)\right\} = \frac{u_0i_0}{i_0^2 + \sigma_i^2} = \frac{R_0}{1 + \sigma_i^2/i_0^2} \quad (2.18)
\]

where the first equality uses \(\mathbb{E}\{u(k)i(k)\} = \mathbb{E}\{u(k)\}\mathbb{E}\{i(k)\}\) and \(\mathbb{E}\{i^2(k)\} = i_0^2 + \mathbb{E}\{n_i^2(k)\}\). The second equality is obtained by dividing the numerator and denominator by \(i_0^2\). Result (2.18) explains why the least squares estimator in Figure 1.3 converges to a value that is too small. Taking the values \(R_0 = 1\ \Omega\), \(i_0 = 1\ \text{A}\), and \(\sigma_i = 0.5\ \text{A}\) of the simulation, we find the limit value 0.8 \(\Omega\), which is consistent with end value of \(\hat{R}_{LS}(N)\) in Figure 1.3. We conclude that the LS estimate is inconsistent.

Similarly, we find for the maximum likelihood solution (1.5b)

\[
\hat{R}_{ML}(N) \xrightarrow{w.p.1} \frac{1}{N} \sum_{k=1}^{N} \frac{\mathbb{E}\{u(k)\}}{\mathbb{E}\{i(k)\}} = \frac{u_0}{i_0} = R_0 \quad (2.19)
\]

which is consistent with end value of \(\hat{R}_{ML}(N)\) in Figure 1.3. We conclude that the ML estimate is strongly consistent. Note, however, that neither the expected value nor the variance of \(\hat{R}_{ML}(N)\) (1.5b) exist for normally distributed current measurements. Indeed, the sum of normally distributed random variables is normally distributed, and the expected value of the inverse of a normally distributed random variable does not exist. Although the probability
mass of $\hat{R}_{\text{ML}}(N)$ is concentrated around the true value for large values of $N$ (strong consistency), the skirts of the pdf of $\hat{R}_{\text{ML}}(N)$ do not decrease sufficiently fast to zero to ensure the existence of the expected value and the variance of the maximum likelihood estimator.

The asymptotic bias (2.18) of the least squares estimate (1.4b) is due to the squaring of the noisy current. This bias can be removed by subtracting the current noise variance from the denominator of (1.4b). The resulting estimator

$$\hat{R}_{\text{BCLS}}(N) = \frac{1}{N} \sum_{k=1}^{N} \frac{u(k)i(k)}{i^2(k) - \sigma^2_i}$$

(2.20)
is called the bias compensated least squares (BCLS) method. Show as an exercise that $\hat{R}_{\text{BLS}}(N)$ (2.20) is a strongly consistent estimate.

### 2.3.2. Asymptotic Variance

Following the lines of Section 2.2.4, the asymptotic variance of the least squares and maximum likelihood estimates of the resistor value are obtained via a first order Taylor series expansion of (1.4b) and (1.5b) around their limit value. For this purpose, all sums are split into a non-random part and a zero mean random part. Since the relative contribution of the zero mean random part will decrease to zero w.p. 1 (strong convergence of the estimators), a Taylor series expansion w.r.t. the random parts can be made.

For the least squares solution (1.4b) we get

$$\frac{1}{N} \sum_{k=1}^{N} u(k)i(k) = u_0i_0 + w_1(N) = u_0i_0(1 + \frac{w_1(N)}{u_0i_0})$$

(2.21a)

$$\frac{1}{N} \sum_{k=1}^{N} i^2(k) = i_0^2 + \sigma_i^2 + w_2(N) = (i_0^2 + \sigma_i^2)(1 + \frac{w_2(N)}{i_0^2 + \sigma_i^2})$$

(2.21b)

where the zero mean random parts $w_1(N)$ and $w_2(N)$ are given by

$$w_1(N) = \frac{1}{N} \sum_{k=1}^{N} n_u(k)i_0 + n_i(k)u_0 + n_u(k)n_i(k)$$

(2.22a)

$$w_2(N) = \frac{1}{N} \sum_{k=1}^{N} 2i_0n_i(k) + n_i^2(k) - \sigma_i^2$$

(2.22b)

Applying the strong law of large numbers (2.8a) to (2.22) we find

$$\frac{w_1(N)}{u_0i_0} \xrightarrow{\text{w.p.1}} 0$$

(2.23a)

$$\frac{w_2(N)}{i_0^2 + \sigma_i^2} \xrightarrow{\text{w.p.1}} 0$$

(2.23b)
Collecting (1.4b) and (2.21)–(2.23), it follows that, with probability one, a first order Taylor series expansion of \( \hat{R}_{LS}(N) \) can be made

\[
\hat{R}_{LS}(N) = \frac{R_0}{1 + \sigma_i^2/i_0^2} \left( 1 + \frac{w_1(N)}{u_0 i_0} + \frac{w_2(N)}{i_0^2 + \sigma_i^2} \right)
\]

\[
\approx \frac{R_0}{1 + \sigma_i^2/i_0^2} \left( 1 + \frac{w_1(N)}{u_0 i_0} - \frac{w_2(N)}{i_0^2 + \sigma_i^2} \right)
\]

(2.24)

From (2.24) we conclude that

\[
\hat{R}_{LS}(N) - \frac{R_0}{1 + \sigma_i^2/i_0^2} \xrightarrow{w.p.1} \delta_{RLS}(N) = \frac{R_0}{1 + \sigma_i^2/i_0^2} \left( \frac{w_1(N)}{u_0 i_0} - \frac{w_2(N)}{i_0^2 + \sigma_i^2} \right)
\]

(2.25)

The asymptotic variance of \( \hat{R}_{LS}(N) \) is now defined as the variance of \( \delta_{RLS}(N) \)

\[
\text{var}(\hat{R}_{LS}(N)) = \text{var}(\delta_{RLS}(N)) = \frac{R_0^2}{(1 + \sigma_i^2/i_0^2)^2} \text{var}\left( \frac{w_1(N)}{u_0 i_0} - \frac{w_2(N)}{i_0^2 + \sigma_i^2} \right)
\]

(2.26)

\[
= \frac{R_0^2}{N(1 + \sigma_i^2/i_0^2)^2} \left( \frac{\sigma_u^2}{u_0^2} + \frac{\sigma_i^2}{i_0^2} + \frac{\sigma_i^2}{u_0^2 \sigma_i^2} - 2 \frac{\sigma_i^2}{i_0^2 + \sigma_i^2} \right)
\]

(2.27)

where the last equality uses \( \mathbb{E}\{n_i^4(k)\} = 3\sigma_i^4 \) and \( \mathbb{E}\{n_i^3(k)\} = 0 \) for zero mean normally distributed noise \( n_i(k) \) [Exercise: prove (2.27) ]. Note that (2.27) is valid for any current and voltage signal-to-noise ratio and that it decreases to zero as an \( O(N^{-1}) \).

Similarly, we obtain for the maximum likelihood estimator (1.5b)

\[
\frac{1}{N} \sum_{k=1}^{N} u(k) = u_0 + w_1(N) = u_0 (1 + \frac{w_1(N)}{u_0})
\]

(2.28a)

\[
\frac{1}{N} \sum_{k=1}^{N} i(k) = i_0 + w_2(N) = i_0 (1 + \frac{w_2(N)}{i_0})
\]

(2.28b)

where the zero mean random parts \( w_1(N) \) and \( w_2(N) \) are given by

\[
w_1(N) = \frac{1}{N} \sum_{k=1}^{N} n_u(k)
\]

(2.29a)

\[
w_2(N) = \frac{1}{N} \sum_{k=1}^{N} n_i(k)
\]

(2.29b)
Applying the strong law of large numbers (2.8a) to (2.29) we find

\[
\frac{w_1(N)}{u_0} \xrightarrow{w.p.} 0 \quad (2.30a)
\]
\[
\frac{w_2(N)}{i_0} \xrightarrow{w.p.} 0 \quad (2.30b)
\]

Collecting (1.5b) and (2.28)–(2.30), it follows that, with probability one, a first order Taylor series expansion of \( \hat{R}_{\text{ML}}(N) \) can be made

\[
\hat{R}_{\text{ML}}(N) = R_0 \left( 1 + \frac{w_1(N)}{u_0} - \frac{w_2(N)}{i_0} \right) \approx R_0 \left( 1 + \frac{w_1(N)}{u_0} - \frac{w_2(N)}{i_0} \right) \quad (2.31)
\]

From (2.31) we conclude that

\[
\hat{R}_{\text{ML}}(N) - R_0 \xrightarrow{w.p.} \delta_{\hat{R}_{\text{ML}}}(N) = R_0 \left( \frac{w_1(N)}{u_0} - \frac{w_2(N)}{i_0} \right) \quad (2.32)
\]

The asymptotic variance of \( \hat{R}_{\text{ML}}(N) \) is now defined as the variance of \( \delta_{\hat{R}_{\text{ML}}}(N) \)

\[
\text{"var}(\hat{R}_{\text{ML}}(N))" = \text{var}(\delta_{\hat{R}_{\text{ML}}}(N)) = R_0^2 \text{var} \left( \frac{w_1(N)}{u_0} - \frac{w_2(N)}{i_0} \right) \quad (2.33)
\]
\[
= R_0^2 \text{var} \left( \frac{w_1(N)}{u_0} - \frac{w_2(N)}{i_0} \right) \quad (2.34)
\]

Note that (2.34) is valid for any current and voltage signal-to-noise ratio and that it decreases to zero as an \( O(N^{-1}) \).

Choosing \( N = 100, R_0 = 1 \Omega, i_0 = 1 \text{A}, u_0 = 1 \text{V}, and \sigma = \sigma_i/i_0 = \sigma_u/u_0, the asymptotic standard deviations (2.27) and (2.34) of, respectively, the least squares and maximum likelihood estimates, are compared in Figure 2.4 for increasing noise levels \( \sigma \). It can be seen that the least squares estimate has a smaller standard deviation for larger noise levels. However, the mean squared error (MSE) calculated as

\[
\text{"MSE}(\hat{R}(N))" = \text{var}(\hat{R}(N)) + (a.s. \lim_{N \to \infty} \hat{R}(N) - R_0)^2 \quad (2.35)
\]

is always larger for \( \hat{R}_{\text{LS}}(N) \) [the right plot of Figure 2.4 shows the square root of (2.35)]. Finally, the variance of the bias compensated least squares estimate (2.20)

\[
\text{"var}(\hat{R}_{\text{BCLS}}(N))" = \frac{R_0^2}{N} \left( \frac{\sigma_u^2}{u_0^2} + \frac{\sigma_i^2}{i_0^2} + \frac{\sigma_u^2 \sigma_i^2}{u_0^2 i_0^2} + 2 \frac{\sigma_i^4}{i_0^4} \right) \quad (2.36)
\]
Figure 2.4: Comparison of the standard deviation (left) and root mean square error (right) of the maximum likelihood (black lines), the least squares (red dashed lines), and the bias compensated least squares (green dashed lines) estimates of the resistor value for increasing noise level $\sigma$.

[Exercise: prove (2.36)] and its mean squared error (2.35) are also given in Figure 2.4 [green dashed lines]. Obviously, $\hat{R}_{BCLS}(N)$ (2.20) has the largest standard deviation, but its root mean square value is much closer to that of $\hat{R}_{ML}(N)$ than $\hat{R}_{LS}(N)$.

Exercise: set up a Monte-Carlo simulation in Matlab to verify the predicted standard deviations and root mean square errors of the least squares, maximum likelihood, and bias compensated least squares estimators in Figure 2.4 [hint: take a sufficiently large number of Monte-Carlo runs and calculate the sample means and sample variances of the estimates over these runs].

2.3.3. Asymptotic Normality

Since the conditions of the second version of the central limit theorem (2.10) are fulfilled for each of the sums in (2.22) and (2.29), the corresponding $w_1(N)$ and $w_2(N)$ are asymptotically normally distributed. The limit random variables $\delta_{R_{LS}}(N)$ (2.25) and $\delta_{R_{ML}}(N)$ (2.32) are a linear combination of the sums $w_1(N)$ and $w_2(N)$, and, therefore, they are also asymptotically normally distributed with zero mean and variances given by (2.27) and (2.34), respectively. Finally, the estimates $\hat{R}_{LS}(N)$ and $\hat{R}_{ML}(N)$ converge w.p. 1 to, respectively, $\delta_{R_{LS}}(N)$ and $\delta_{R_{ML}}(N)$, and, hence, also in distribution [see Figure 2.1], which proves the asymptotic normality of the least squares (1.4b) and maximum likelihood (1.5b) estimates.

A similar reasoning can held to prove that the bias compensated least squares estimator (2.20) is asymptotically normally distributed with mean
value $R_0$ and variance $(2.36)$.

The important consequence of the asymptotic normality is that tight uncertainty bound with a given confidence level can be constructed. For example, $\hat{R}(N) \pm \text{std}(\hat{R}(N))$ and $\hat{R}(N) \pm 2\text{std}(\hat{R}(N))$ are confidence intervals with a confidence level of, respectively, 68% and 95%.

2.3.4. Asymptotic Efficiency

To calculate the Cramér-Rao lower bound (2.15) for the resistor estimation problem, we have to construct the probability density function (pdf) of the measurements $z$

$$z = [u(1) \ u(2) \ \ldots \ u(N) \ i(1) \ i(2) \ \ldots \ i(N)]^T \quad (2.37)$$

Since all current $i(k)$ and voltage $u(k)$ measurements, $k = 1, 2, \ldots, N$, are independently distributed, the pdf of $z$ equals the product of the pdfs of each current and voltage measurement separately

$$f_{z|\theta}(z|\theta) = \frac{1}{(2\pi \sigma_u \sigma_i)^N} e^{-\frac{1}{2} \sum_{k=1}^{N} \frac{(u(k) - R_0 i_0)^2}{2\sigma_u^2} - \sum_{k=1}^{N} \frac{(i(k) - i_0)^2}{2\sigma_i^2}}$$

Within a parameter independent constant, the natural logarithm of (2.38) equals

$$\log f_{z|\theta}(z|\theta) = -\sum_{k=1}^{N} \frac{(u(k) - R_0 i_0)^2}{2\sigma_u^2} - \sum_{k=1}^{N} \frac{(i(k) - i_0)^2}{2\sigma_i^2} + (u(k) - R_0 i_0)^2 \sigma_u^2 + (i(k) - i_0)^2 \sigma_i^2$$

from which we derive its second order derivative w.r.t. $\theta_0 = [R_0 \ i_0]^T$

$$-\frac{\partial^2 f_{z|\theta}(z|\theta)}{\partial \theta_0^2} = \begin{bmatrix} \sum_{k=1}^{N} \frac{i_0^2}{\sigma_i^2} - \sum_{k=1}^{N} \frac{u(k) - 2R_0 i_0}{\sigma_u^2} & -\sum_{k=1}^{N} \frac{u(k) - 2R_0 i_0}{\sigma_u^2} \\ -\sum_{k=1}^{N} \frac{u(k) - 2R_0 i_0}{\sigma_u^2} & \sum_{k=1}^{N} \frac{R_0^2}{\sigma_u^2} + \frac{1}{\sigma_i^2} \end{bmatrix} \quad (2.40)$$

The expected value of (2.40) gives the Fisher information matrix (2.16b)

$$\text{Fi}(\theta_0) = \begin{bmatrix} N \frac{\partial^2}{\sigma_i^2} & N \frac{R_{0\theta_0}}{\sigma_u^2} \\ N \frac{R_{0\theta_0}}{\sigma_u^2} & N \frac{R_{0\theta_0}^2 + \sigma_u^2}{\sigma_i^2} \end{bmatrix} \quad (2.41)$$

From the inverse of (2.41),

$$\text{Fi}^{-1}(\theta_0) = \frac{1}{N} \begin{bmatrix} \frac{R_0^2 \sigma_i^2 + \sigma_u^2}{\sigma_i^2} & \frac{-R_0 \sigma_i^2}{\sigma_i^2} \\ \frac{-R_0 \sigma_i^2}{\sigma_i^2} & \frac{-\sigma_i^2}{\sigma_i^2} \end{bmatrix} \quad (2.42)$$
and using $u_0 = R_0 i_0$, we find the Cramér-Rao lower bound for estimating the resistor value from noisy DC current and voltage measurements

$$\text{var}(\hat{R}(N)) \geq \frac{R_0^2}{N} \left( \frac{\sigma_i^2}{i_0^2} + \frac{\sigma_u^2}{u_0^2} \right)$$

Comparing (2.34) and (2.36) to (2.43), it can be concluded that the maximum likelihood estimator (1.5b) is asymptotically efficient, but not the bias compensated least squares (2.20).

### 2.3.5. Robustness

When constructing the least squares solution (1.4), we silently assumed that the DC current is exactly known and that the DC voltage measurements are independent and normally distributed. Under these conditions, the least squares estimate is strongly consistent [(2.18) with $\sigma_i = 0$], asymptotically normally distributed, and asymptotically efficient [compare (2.27) to (2.43) for the case $\sigma_i = 0$]. The following observations can be made:

1. **Non-Gaussian iid voltage measurements with finite variance.** The least squares estimator is still consistent [the strong law of larger numbers (2.8a) is valid], asymptotically normally distributed [the first version of the central limit theorem (2.10) is applicable], with an asymptotic variance that can be calculated from (2.27) with $\sigma_i = 0$ [the first order Taylor series expansion does not use the distribution function of the measurements]. Only the asymptotic efficiency is no longer valid and, hence, not robust w.r.t. the Gaussian measurement assumption.

2. **Correlated voltage measurements (mixing of order infinity).** The least squares estimator is still consistent [the law of larger numbers (2.8c) is valid], asymptotically normally distributed [the third version of the central limit theorem (2.10) is applicable], with an asymptotic variance that can be calculated from (2.26) [the first order Taylor series expansion does not use the distribution function of the measurements]. Only the particular variance expression (2.27) and the asymptotic efficiency are no longer valid and, hence, not robust w.r.t. the Gaussian measurement assumption.

3. **Noisy current measurements.** The least squares estimator is no longer consistent [see (2.18)], nor asymptotically efficient. However, the estimator is still asymptotically normally distributed with asymptotic variance
We conclude that the consistency and asymptotic efficiency properties of (1.4) are not robust w.r.t. the assumption that the DC current is known exactly.

To construct the maximum likelihood estimator (1.4) it has silently been assumed that the current and voltage measurements are mutually independent, and independent and identically normally distributed random variables. What happens with the asymptotic properties if one or more of these assumptions are violated is discussed below:

1. **Non-Gaussian iid current and voltage measurements with finite variances.** The maximum likelihood estimator is still consistent [the strong law of larger numbers (2.8a) is valid], asymptotically normally distributed [the first version of the central limit theorem (2.10) is applicable], with an asymptotic variance given by (2.34) [the first order Taylor series expansion does not use the distribution function of the measurements]. Only the asymptotic efficiency is no longer valid and, hence, not robust w.r.t. the Gaussian current and voltage measurement assumption.

2. **Correlated measurements (mixing of order infinity)** The maximum likelihood estimator is still consistent [the law of larger numbers (2.8c) is valid], asymptotically normally distributed [the third version of the central limit theorem (2.10) is applicable], with an asymptotic variance that can be calculated from (2.33) [the first order Taylor series expansion does not use the distribution function of the measurements]. Only the particular variance expression (2.34) and the asymptotic efficiency are no longer valid and, hence, not robust w.r.t. the Gaussian measurement assumption.

**2.4. Analysis of Estimators Defined Implicitly as the Minimizing Argument of a Cost Function**

In this section we consider the case where the estimator $\hat{\theta}(z) \in \mathbb{R}^{n_0 \times 1}$ is defined as the minimizing argument of a cost function $V(\theta, z)$ that cannot be minimized analytically w.r.t. $\theta$

$$\hat{\theta}(z) = \arg \min_{\theta} V(\theta, z)$$ (2.44)

where $z \in \mathbb{R}^{N \times 1}$ is the vector of the noisy data. We analyze the properties of (2.44) for $N \to \infty$. Therefore, to keep the asymptotic cost function finite,
$V(\theta, z)$ is scaled by $N$

$$V_N(\theta, z) = \frac{1}{N} V(\theta, z) \quad (2.45)$$

### 2.4.1. Consistency

Necessary conditions for consistency are:

1. The expected value of the cost function

$$V_N(\theta) = \mathbb{E}\{V_N(\theta, z)\} \quad (2.46)$$

is minimal in the true model parameters $\theta_0$

$$\theta_0 = \arg \min_\theta V_N(\theta) \quad (2.47)$$

Note that condition (2.47) can be relaxed to be true for the limit of the expected value of the cost function

$$V_*(\theta) = \lim_{N \to \infty} V_N(\theta) \quad (2.48)$$

2. The cost function converges in stochastic sense to its expected value (application of one of the laws of large numbers), for example, mean square convergence,

$$\lim_{N \to \infty} \text{var}(V_N(\theta, z)) = 0 \quad (2.49)$$

*Exercise: prove the equivalence in (2.49).*

To conclude consistency from (2.47) and (2.49), two technical conditions should be satisfied (Söderström, 1974). First, the convergence (2.49) should be uniform w.r.t. $\theta$ in a closed and bounded neighborhood of $\theta_0$. Next, the cost function (2.45) and its derivative w.r.t. $\theta$ should be continuous in that closed and bounded neighborhood of $\theta_0$.

Figure 2.5 illustrates what happens if the convergence (2.49) of the cost function is not uniform. In that case (2.49) does not imply convergence of the global minimizer of $V_N(\theta, z)$ (2.45) to the global minimizer of $V_*(\theta)$ (2.48).

### 2.4.2. Asymptotic Covariance

To obtain an asymptotic expression for the difference between the estimate $\hat{\theta}(z)$ and the true value $\theta_0$, a first order Taylor series expansion with remainder of the derivative of the cost function w.r.t. $\theta$ is made

$$V_N^T(\hat{\theta}(z), z) = V_N^T(\theta_0, z) + V_N''(\theta_1, z)(\hat{\theta}(z) - \theta_0) \quad (2.50)$$
where $\theta_1 = t\hat{\theta}(z) + (1 - t)\theta_0$ with $t \in [0,1]$ (Kaplan, 1993). Note that (2.50) is exact without any approximation. The following observations can be made:

1. By definition of the estimate $\hat{\theta}(z)$
   \[ V_N^T(\hat{\theta}(z), z) = 0 \] (2.51)

2. $\hat{\theta}(z)$ converges in stochastic sense to $\theta_0$ (consistency) and, hence,
   \[ \theta_1 \xrightarrow{{\text{stoch. sense}}} N \rightarrow \infty t\theta_0 + (1 - t)\theta_0 = \theta_0 \] (2.52)

3. The convergence of the second order derivative of the cost function to its expected value
   \[ V_N''(\theta, z) \xrightarrow{{\text{stoch. sense}}} N \rightarrow \infty V_N''(\theta) \] (2.53)
   follows from (2.49).

4. The expected value of the first term in the right hand side of (2.50) is zero
   \[ \mathbb{E}\{V_N^T(\theta_0, z)\} = V_N^T(\theta_0) = 0 \] (2.54)
   where the last equality uses (2.47).

Collecting (2.50)–(2.54), we find
\[ \hat{\theta}(z) - \theta_0 \xrightarrow{{\text{stoch. sense}}} N \rightarrow \infty \delta_\theta(z) = -V_N''^{-1}(\theta_0)V_N^T(\theta_0, z) \] (2.55)
where $\mathbb{E}\{\delta_\theta(z)\} = 0$ [see (2.54)]. Hence, the asymptotic covariance of $\hat{\theta}(z)$ is given by

$$
\text{“Cov}(\hat{\theta}(z))" = \text{Cov}(\delta_\theta(z)) = \left(\frac{1}{N}\right) \mathbb{E}\{V_N'(\theta_0, z)V_N'N(\theta_0, z)\}V_N''^{-1}(\theta_0) \quad (2.56)
$$

2.4.3. Asymptotic Efficiency

Comparing the asymptotic covariance matrix (2.56) to the Cramér-Rao lower bound (2.15) establishes the asymptotic (in)efficiency of the estimator.

2.4.4. Asymptotic Normality

The asymptotic distribution of $\hat{\theta}(z)$ is set by the asymptotic distribution of $\delta_\theta(z)$ [see (2.55)]. Indeed, strong or weak convergence or convergence in mean square sense implies convergence in distribution [see Figure 2.1 on page 14]. Since the inverse matrix in (2.55) is deterministic, we only have to verify whether, according to one of the versions of the central limit theorem (2.10), the $n_\theta \times 1$ vector $V_N'(\theta_0, z)$ is asymptotically ($N \to \infty$) normally distributed.

2.4.5. Robustness

In general, the particular distribution function of $z$ is not needed to establish the stochastic convergence (2.49) of the cost function. Also the derivation of the limit random variable $\delta_\theta(z)$ (2.55), the general expression of the asymptotic covariance (2.56), and the asymptotic normality do not make use of the particular pdf of $z$. Hence, all these properties are robust w.r.t. the assumed distribution function of $z$ for constructing the cost function $V(\theta, z)$.

2.5. Asymptotic Covariance of a Function of a Consistent Estimator

Assume that $\hat{\theta}(N) \in \mathbb{R}^{n_\theta \times 1}$ converges for $N \to \infty$ w.p. 1 to $\theta_0$ with asymptotic covariance matrix $C_\theta$. In this section we study the asymptotic properties of a continuous and twice differentiable function $\hat{\psi}(N) = f(\hat{\theta}(N)) \in \mathbb{R}^{n_\psi \times 1}$ of $\hat{\theta}(N)$.

Using the interchangeability of the almost sure limit (2.1) and a continuous function, we find

$$
a.s. \lim_{N \to \infty} \hat{\psi}(N) = f(a.s. \lim_{N \to \infty} \hat{\theta}(N)) = f(\theta_0) = \psi_0 \quad (2.57)
$$
which shows the consistency of $\hat{\psi}(N)$.

A second order Taylor series expansion with remainder of $f(\hat{\theta}(N))$ at $\theta_0$, gives

$$
\hat{\psi}(N) = \psi_0 + f'(\theta_0)(\hat{\theta}(N) - \theta_0) + \frac{1}{2}(\hat{\theta}(N) - \theta_0)^T f''(\theta_1)(\hat{\theta}(N) - \theta_0) \quad (2.58)
$$

where $\theta_1 = t\theta_0 + (1 - t)\hat{\theta}(N)$, with $t \in [0, 1]$ (Kaplan, 1993). Since $\hat{\theta}(N)$ is a strongly consistent estimate of $\theta_0$ the third term in (2.58) converges faster to zero than the second term. Hence, it follows from (2.58) that

$$
\hat{\psi}(N) - \psi_0 \xrightarrow{\text{w.p.1}} \frac{\delta_\psi(N)}{N \to \infty} = f'(\theta_0)(\hat{\theta}(N) - \theta_0) \quad (2.59)
$$

from which we get the asymptotic covariance of $\hat{\psi}(N)$

$$
\text{Cov}(\hat{\psi}(N)) = \text{Cov}(\delta_\psi(N)) = f'(\theta_0)C_{\theta}f'^T(\theta_0) \quad (2.60)
$$
2.6. Appendix: Proof of the Cramér-Rao Lower Bound

The proof of the Cramér-Rao lower bound is based on the Cauchy-Schwarz inequality for random vectors. First, we derive this inequality and, next, the lower bound (2.15) follows as a special case.

2.6.1. Cauchy-Schwartz Inequality

Let $U \in \mathbb{R}^n$ and $V \in \mathbb{R}^k$ be random vectors. Define the $n \times n$ matrix $M$ as

$$ M = \mathbb{E}\{(U - \Gamma V)(U - \Gamma V)^T\} \quad (2.61) $$

where $\Gamma \in \mathbb{R}^{n \times k}$ is a deterministic matrix. By construction, $M$ (2.61) is positive semi-definite

$$ M \succeq 0 \quad (2.62) $$

Combining (2.61) and (2.62) with the choice

$$ \Gamma = \mathbb{E}\{UV^T\} \mathbb{E}\{VV^T\}^{-1} \quad (2.63) $$

we find

$$ \mathbb{E}\{UU^T\} \succeq \mathbb{E}\{UV^T\} \mathbb{E}\{VV^T\}^{-1} \mathbb{E}\{VU^T\} \quad (2.64) $$

which is the Cauchy-Schwartz inequality.

2.6.2. Cramér-Rao Lower Bound

We choose

$$ U = \hat{\theta}(z) - \theta_0 \quad (2.65) $$

$$ V^T = \frac{\partial \log f_{z|\theta_0}(z|\theta_0)}{\partial \theta_0} \quad (2.66) $$

in the Cauchy-Schwartz inequality (2.64). Taking into account that

$$ \mathbb{E}\{\hat{\theta}(z)\} = \theta_0 \quad (2.67a) $$

$$ \mathbb{E}\left\{\frac{\partial \log f_{z|\theta_0}(z|\theta_0)}{\partial \theta_0}\right\} = \frac{\partial}{\partial \theta_0} \int_Z f_{z|\theta_0}(z|\theta_0) dz = \frac{\partial 1}{\partial \theta_0} = 0 \quad (2.67b) $$

$$ \mathbb{E}\left\{\theta_0 \frac{\partial \log f_{z|\theta_0}(z|\theta_0)}{\partial \theta_0}\right\} = \theta_0 \mathbb{E}\left\{\frac{\partial \log f_{z|\theta_0}(z|\theta_0)}{\partial \theta_0}\right\} = 0 \quad (2.67c) $$

$$ \mathbb{E}\left\{\hat{\theta}(z) \frac{\partial \log f_{z|\theta_0}(z|\theta_0)}{\partial \theta_0}\right\} = \frac{\partial}{\partial \theta_0} \int_Z \hat{\theta}(z)f_{z|\theta_0}(z|\theta_0) dz = \frac{\partial \theta_0}{\partial \theta_0} = I_{\theta_0} \quad (2.67d) $$
where Fi is defined in (2.16a). Note that first equalities of (2.67b) and (2.67d) use the assumption that the domain $Z$ of $f_{z|\theta_0}(z|\theta_0)$ does not depend on $\theta_0$.

Combining (2.64) with (2.68) proves (2.15) and (2.16a).

Differentiation (2.67b) w.r.t. $\theta_0$ we find

$$
\frac{\partial}{\partial \theta_0} \int_Z \frac{\partial \log f_{z|\theta_0}(z|\theta_0)}{\partial \theta_0} \log f_{z|\theta_0}(z|\theta_0) dz = 0
$$

$$
\Rightarrow \mathbb{E} \left\{ \frac{\partial^2 \log f_{z|\theta_0}(z|\theta_0)}{\partial \theta_0^2} \right\} + \mathbb{E} \left\{ \left( \frac{\partial \log f_{z|\theta_0}(z|\theta_0)}{\partial \theta_0} \right)^T \left( \frac{\partial \log f_{z|\theta_0}(z|\theta_0)}{\partial \theta_0} \right) \right\} = 0
$$

$$
\Rightarrow \mathbb{E} \left\{ \frac{\partial^2 \log f_{z|\theta_0}(z|\theta_0)}{\partial \theta_0^2} \right\} = -\mathbb{E} \left\{ \left( \frac{\partial \log f_{z|\theta_0}(z|\theta_0)}{\partial \theta_0} \right)^T \left( \frac{\partial \log f_{z|\theta_0}(z|\theta_0)}{\partial \theta_0} \right) \right\}
$$

Combining (2.16a) with (2.69) finally proves (2.16b).
Chapter 3
Linear Least Squares

Abstract: Quite a number of modeling problems can be formulated or approximated as a linear regression. This chapter gives an in depth analysis of linear regression problems from a stochastic as well as a numerical point of view.

Learning Objectives:

- Derivation of the finite sample and the asymptotic properties of (weighted) linear least squares estimators.
- Numerical stable calculation of (weighted) linear least squares estimators – understanding the numerical issue of forming the matrix product $H^T H$.
- Understanding the danger of extrapolation.
- Elimination of the bias induced by the noise on the regression matrix.
- Regularization: a tool to make a bias-variance trade-off in case the number of parameters is not much smaller than the amount of data.
- Handling outliers in measurements.

3.1. Problem Statement

In this chapter we consider models that are linear in the parameters $\theta_0 \in \mathbb{R}^{n_0 \times 1}$. Assuming that noisy observations $y \in \mathbb{R}^{N \times 1}$ are available, the model can be written as

$$y = H\theta_0 + v \tag{3.1}$$

where $H \in \mathbb{R}^{N \times n_0}$ is the regression matrix and with $v \in \mathbb{R}^{N \times 1}$ the zero mean disturbing noise, $\mathbb{E}\{v\} = 0$, with covariance matrix $\text{Cov}(v) = C_v$. 

35
In a classical setting, the regression matrix $H$ is assumed to be independent of the disturbing noise [Section 3.2], and the model parameters are obtained by minimizing
\[
V_{LS}(\theta, y) = \frac{1}{2} (y - H\theta)^T (y - H\theta) \tag{3.2}
\]
w.r.t. $\theta$. If the regression matrix depends on $v$, then the nice properties of the linear least squares estimator are – in general – lost. Section 3.3 presents two methods that can cope with this issue.

In some applications such as, for example, the estimation of the impulse response coefficients of a discrete-time system from known input, noisy output measurements, the number $n_\theta$ of parameters to be estimated is not much smaller than the amount of data $N$. This results in highly variable estimates [see, for example, Section 1.2.2 on page 6]. A solution to this problem is regularization that makes a bias-variance trade-off [Section 3.4].

Sometimes measurements are corrupted by outliers (large unlikely values) that strongly affect the least squares cost function (3.2) and, hence, also the linear least squares estimate. Section 3.5 explains how to deal with this issue.

### 3.2. The Linear Least Squares Estimator – Noiseless Regression Matrix

Minimization of (3.2) w.r.t. $\theta$ gives the linear least squares estimate $\hat{\theta}_{LS}(y)$, and the corresponding estimate $\hat{y}_{LS}$
\[
\hat{\theta}_{LS}(y) = (H^T H)^{-1} H^T y \tag{3.3a}
\]
\[
\hat{y}_{LS} = H(H^T H)^{-1} H^T y \tag{3.3b}
\]
First, we discuss the stochastic properties of (3.3a) [Section 3.2.1]. It includes the asymptotic ($N \to \infty$) as well as the finite sample behavior. Next, we discuss the numerical issues that can occur when calculating (3.3), and we propose some solutions to this problem [Section 3.2.2]. The performance of these methods are illustrated on a polynomial curve fitting problem [Section 3.2.3]. Finally, a weighting is added to the cost function (3.2), aiming to reduce the covariance of the estimate [Section 3.2.4]. It is shown which weighting minimizes the covariance of the estimate.

#### 3.2.1. Stochastic Properties

The linear least squares estimator (3.3a) is unbiased
\[
\mathbb{E}\{\hat{\theta}_{LS}(y)\} = \theta_0 + (H^T H)^{-1} H^T \mathbb{E}\{v\} = \theta_0 \tag{3.4}
\]
where the first equality uses (3.1), and the second equality the zero mean property of the disturbing noise \( v \). Hence, the covariance matrix of (3.3a) equals

\[
\text{Cov}(\hat{\theta}_{\text{LS}}(y)) = (H^T H)^{-1}H^T C_v H (H^T H)^{-1}
\]

(3.5)

where \( C_v = \text{Cov}(v) \). Note that (3.4) and (3.5) are independent of the particular distribution function of \( v \), and are valid for finite amount of data \( N \).

Assuming that the noise \( v \) is normally distributed, the pdf of the measurements \( y \), given the true model parameters \( \theta_0 \), is given by

\[
f_{y|\theta_0}(y|\theta_0) = \frac{1}{\sqrt{(2\pi)^N \det C_v}} e^{-\frac{1}{2} (y-H\theta_0)^T C_v^{-1} (y-H\theta_0)}
\]

(3.6)

Using (3.6), the Fisher information matrix \( Fi(\theta_0) \) (2.16b) is readily found

\[
Fi(\theta_0) = H^T C_v^{-1} H
\]

(3.7)

which results in the following Cramér-Rao lower bound

\[
\text{Cov}(\hat{\theta}(y)) \geq (H^T C_v^{-1} H)^{-1}
\]

(3.8)

Comparing (3.5) with the lower bound (3.8), it follows that \( \text{Cov}(\hat{\theta}_{\text{LS}}(y)) > Fi^{-1}(\theta_0) \) [proof: see Appendix 3.7 on 60] and, hence, the linear least squares estimator (3.3a) is inefficient. It suggests that an estimator can be constructed with a smaller covariance [see Section 3.2.4].

To analyze the stochastic convergence \( (N \to \infty) \) of the linear least squares estimator, we combine (3.1) and (3.3a), which gives

\[
\hat{\theta}_{\text{LS}}(y) = \theta_0 + (H^T H)^{-1} H^T v
\]

(3.9)

It follows that \( \hat{\theta}_{\text{LS}}(y) \) is a consistent estimator if and only if

\[
\delta_\theta(v) = (H^T H)^{-1} H^T v \overset{\text{stoch. sense}}{\underset{N \to \infty}{\to}} 0
\]

(3.10)

is satisfied. This is the case under the following technical conditions on the regression matrix \( H \) and the disturbing noise \( v \)

1. Increasing the number of data point \( N \) should add information about the model parameters. This is mathematically translated as

\[
\lim_{N \to \infty} \frac{1}{N} H^T H \text{ is of full rank}
\]

(3.11)
2. The correlation of the noise over the measurements should decrease sufficiently fast to zero. This is mathematically translated as a mixing condition of order \( P \geq 2 \). It implies that the off-diagonal elements of \( C_v \) decrease sufficiently fast to zero as the distance to the main diagonal increases.

3. The 1-norm of the scaled regression matrix \( N^{-1}H \) should be finite

\[
\frac{1}{N}||H||_1 \leq c_1 < \infty \Leftrightarrow \frac{1}{N} \max_{l=1,\ldots,n_\theta} \sum_{k=1}^{N} |H[k,l]| \leq c_1 < \infty \quad (3.12)
\]

for \( N = 1, 2, \ldots, \infty \) and with \( c_1 \) a constant independent of \( N \).

Indeed, if conditions 1–3 are fulfilled with \( P = 2 \), then, \( \delta_\theta(v) \) (3.10) is mixing of order 2 [proof: see Pintelon and Schoukens, 2012, p. 575, Lemma 16.5]. Hence, the law of large numbers (2.8c) is applicable to \( \delta_\theta(v) \) such that

\[
\begin{align*}
\lim_{N \to \infty} \delta_\theta(v) &= 0 \\
\text{asymptotically normally distributed, with mean value } \theta_0 \text{ and covariance matrix } (3.5).
\end{align*}
\]

which proves (3.10).

If technical conditions 1–3 are fulfilled with \( P = \infty \), then, \( \delta_\theta(v) \) (3.10) is mixing of order infinity [see Pintelon and Schoukens, 2012, p. 575, Lemma 16.5]. Hence, the third version of the central limit theorem in Section 2.2.5 on page 16 can be applied to \( \delta_\theta(v) \), which shows that \( \hat{\theta}_{\text{LS}}(y) \) is asymptotically normally distributed, with mean value \( \theta_0 \) and covariance matrix (3.5).

We conclude that the linear least squares estimator is – in general – an unbiased, inefficient, asymptotically normally distributed and consistent estimator.

### 3.2.2. Numerical Stable Calculation

Calculating the linear least squares estimate straightforwardly via (3.3a), might result in significant numerical problems. First, we reveal the key issue via a singular value decomposition (SVD) of the regression matrix \( H \). Next, we propose a few numerical stable implementations of (3.3a).

The singular value decomposition of the regression matrix \( H \in \mathbb{R}^{N \times n_\theta} \) equals (Golub and Van Loan, 1996)

\[
H = U \Sigma V^T \quad (3.14)
\]

where \( U \in \mathbb{R}^{N \times n_\theta} \) is an orthogonal matrix \( (U^T U = I_{n_\theta}, \text{ with } I_{n_\theta} \text{ the } n_\theta \times n_\theta \text{ identity matrix}) \), \( \Sigma \in \mathbb{R}^{n_\theta \times n_\theta} \) a diagonal matrix containing the singular values.
\[ \sigma_k(H) \geq 0, \ k = 1, 2, \ldots, n_\theta, \] and \( V \in \mathbb{R}^{n_\theta \times n_\theta} \) a square orthogonal matrix \( (V^{-1} = V^T) \). The condition number \( \kappa(H) \) of the regression matrix \( H \)

\[ \kappa(H) = \frac{\max_k \sigma_k(H)}{\min_k \sigma_k(H)} \tag{3.15} \]

gives a lower bound on the number of significant digits \( n_{\text{sd}} \) needed to represent accurately \( H \)

\[ n_{\text{sd}} \gg \log_{10} \kappa(H) \tag{3.16} \]

If (3.16) is not fulfilled, then the matrix is said to be ill-conditioned. Using (3.14), we find for \( H^T H \)

\[ H^T H = VSU^T U V^T = V\Sigma^2 V^T \tag{3.17} \]

It follows that the condition number of \( H^T H \) is the square of that of \( H \)

\[ \kappa(H^T H) = \kappa^2(H) \tag{3.18} \]

Hence, if we form explicitly the matrix product \( H^T H \) in (3.3a), we need at least twice as much significant digits as in (3.16)

\[ n_{\text{sd}} \gg 2 \log_{10} \kappa(H) \tag{3.19} \]

For example, if the condition number of \( H \) is \( 10^8 \), then it can still be accurately represented within the \( n_{\text{sd}} = 15 \) significant digits of Matlab. However, this is no longer true for \( H^T H \), because its condition number is equal to \( 10^{16} \).

By means of the singular value decomposition (3.14), we can calculate the linear least squares estimate (3.3a) without forming explicitly the matrix product \( H^T H \). To prove this, \( (H^T H)^{-1} H^T \) is rewritten as

\[ (H^T H)^{-1} H^T = (V\Sigma^2 V^T)^{-1} V\Sigma U^T = V\Sigma^{-1} U^T \tag{3.20} \]

where the first equality uses \( U^T U = I_{n_\theta} \), and the second equality \( V^{-1} = V^T \).

Combining (3.3) with (3.20) we finally get

\[ \hat{\theta}_{\text{LS}}(y) = V\Sigma^{-1} U^T y \tag{3.21a} \]
\[ \hat{y}_{\text{LS}} = U U^T y \tag{3.21b} \]

where \( V\Sigma^{-1} U^T = H^+ \) is called the pseudo-inverse of \( H \) (Ben-Israel and Greville, 1974).

An alternative way consists in using the QR-factorization of \( H \) (Golub and Van Loan, 1996)

\[ H = QR \tag{3.22} \]
where $Q \in \mathbb{R}^{N \times n_\theta}$ is an orthogonal matrix ($Q^T Q = I_{n_\theta}$) and $R \in \mathbb{R}^{n_\theta \times n_\theta}$ an upper triangular matrix. Substituting (3.22) in $(H^T H)^{-1} H^T$, we find

$$(H^T H)^{-1} H^T = (R^T R)^{-1} R^T Q^T = R^{-1} Q^T$$

(3.23)

where the first equality uses $Q^T Q = I_{n_\theta}$. Combining (3.3) with (3.23) we finally get

$$\hat{\theta}_{LS}(y) = R^{-1} Q^T y$$

(3.24a)

$$\hat{y}_{LS} = QQ^T y$$

(3.24b)

In practice, the inverse matrix $R^{-1}$ is not calculated, but instead the set of equations $R\hat{\theta}_{LS}(y) = Q^T y$ is solved via back substitution. Compared with the SVD solution (3.21a), the QR method (3.24a) is faster but numerically less stable.

Often, the SVD (3.21a) or QR (3.24a) solutions are not sufficient for guaranteeing a good numerical conditioning of the problem. This is due to the fact that the model parameters mostly have different physical units. Changing the units of the model parameters can drastically change the dynamic range of the numbers in the regression matrix and, hence, also its condition number. To avoid mixing of physical units, and in order to minimize the dynamic range of the numbers, each column of the regression matrix $H$ in (3.1) is scaled by its 2-norm

$$y = HS^{-1}S\theta_0 + v \quad \text{with} \quad S_{[k,k]} = \sqrt{\sum_{l=1}^{N} H_{[l,k]}^2}, \quad k = 1, 2, \ldots, n_\theta$$

(3.25)

Next, the SVD or QR factorization of the scaled regression matrix is calculated

$$HS^{-1} = U_S \Sigma_S V_S^T$$

(3.26a)

$$HS^{-1} = Q_S R_S$$

(3.26b)

giving the estimate

$$\hat{\psi}(y) = V_S \Sigma_S^{-1} U_S y$$

(3.27a)

$$\hat{\psi}(y) = R_S^{-1} Q_S^T y$$

(3.27b)

of $\psi_0 = S\theta_0$. Finally, the linear least squares estimates of $\theta$ and $y$ are obtained from (3.26) and (3.27) as

$$\hat{\theta}_{LS}(y) = S^{-1} \hat{\psi}(y)$$

(3.28a)

$$\hat{y}_{LS} = U_S U_S^T y$$

(3.28b)

$$\hat{y}_{LS} = Q_S Q_S^T y$$

(3.28c)
where the last two equations use $\hat{y}_{LS} = (HS^{-1})\hat{\psi}(y)$.

Sometimes, for very complex problems, the combination of scaling the regression matrix and calculating the solution via SVD or QR factorization is still not good enough for satisfying (3.16). In that case, one should replace the basis functions used in (3.1), by either standard orthogonal functions, or orthogonal functions obtained via a Gram-Smidt orthogonalisation procedure that results in a regression matrix with condition number equal to one [see Van Barel and Bultheel, 1992 and Pintelon and Schoukens, 2012, page 558, Section 15.11]. The latter is the best one can do.

3.2.3. Example: Polynomial Curve Fitting

The function $y = \arctan(u)$ is approximated over the interval $[0, 5]$ by a polynomial of order $n_\theta - 1$ in powers of $u$ or powers of $2u/5 - 1$

$$y(u, \theta) = \sum_{r=1}^{n_\theta} \theta_{[r]} u^{r-1}$$

(3.29a)

$$y(u, \theta) = \sum_{r=1}^{n_\theta} \theta_{[r]} \left(\frac{2}{5}u - 1\right)^{r-1}$$

(3.29b)

While (3.29a) and (3.29b) are mathematically equivalent, their numerical properties are quite different as will be shown in the simulation results.

Starting from $N = 200$ equidistantly distributed points in the interval $[0, 5]$

$$u(k) = \frac{5}{N-1}(k-1) \quad \text{with} \quad k = 1, 2, \ldots, N$$

(3.30)

the coefficients $\theta$ of the polynomials (3.29) of degree $n_\theta - 1 = 30$ are estimated in least squares sense (3.2) from the noiseless arctan function [$v = 0$ in (3.1)].

The regression matrix $H$ in (3.1) corresponding to (3.29) equals

$$H_{[k,r]} = u^{r-1}(k)$$

(3.31a)

$$H_{[k,r]} = \left(\frac{2}{5}u(k) - 1\right)^{r-1}$$

(3.31b)

with $k = 1, 2, \ldots, N$ and $r = 1, 2, \ldots, n_\theta$, and the linear least squares solution [minimizer (3.2)] is calculated for the two models (3.29) in three different ways:

1. Direct application of (3.3).
2. Use of the SVD of the regression matrix (3.21).
3. Via the SVD of the scaled regression matrix (3.28).
Figure 3.1: Curve fitting of the arctan function [left] using two different polynomial basis functions of order $n_y - 1 = 30$ and three different numerical methods for calculating the linear least squares solution [right]. Solution (3.3b) for polynomials (3.29a) [cyan] and (3.29b) [magenta]. Solution (3.21b) for polynomials (3.29a) [black] and (3.29b) [red]. Solution (3.28b) for polynomials (3.29a) [green] and (3.29b) [blue, on top of red].

Figure 3.1 shows the magnitude of the relative approximation error in dB

$$\frac{\hat{y}_{L.S.}[k] - \arctan(u(k))}{\arctan(u(k))}$$

(3.32)

for the two polynomial models and the three solution methods, and Table 3.1 the corresponding condition numbers. The following observations can be made:

1. Direct application of (3.3) results in large approximation errors for both models [cyan and magenta curves], which can be explained by the very large condition numbers \( \gg 10^{15} \) of \( H^T H \).

2. Using the singular value decomposition of the (scaled) regression matrix, the relative approximation errors are smaller than -100 dB [black and green curves] and -150 dB [blue and red curves] for, respectively, polynomials models (3.29a) and (3.29b). Hence, centering the argument of the basis polynomials around the mean of the interval [0,5] and scaling with the mean value, \((u - 2.5)/2.5\), results in a gain of 50 dB in the relative approximation error. This can also be explained by the much smaller condition numbers of the SVD$_2$ and SVD$_{sc,2}$ solutions based on model (3.29b), compared with SVD$_1$ and SVD$_{sc,1}$ based on model (3.29a) [see Table 3.1].

3. Scaling of the regression matrix (3.25) improves the condition number: compare the condition number of SVD$_{sc,i}$ with that of SVD$_i$ for \( i = 1, 2 \).
Table 3.1: Condition number of the three least squares solutions LS (3.3), SVD (3.21) and SVD_{sc} (3.28) for the polynomial models (3.29a) [subscript 1] and (3.29b) [subscript 2].

<table>
<thead>
<tr>
<th></th>
<th>LS_1</th>
<th>SVD_1</th>
<th>SVD_{sc,1}</th>
<th>LS_2</th>
<th>SVD_2</th>
<th>SVD_{sc,2}</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>3.5 × 10^{32}</td>
<td>1.9 × 10^{16}</td>
<td>1.3 × 10^{16}</td>
<td>1.2 × 10^{22}</td>
<td>1.1 × 10^{11}</td>
<td>5.8 × 10^{10}</td>
</tr>
</tbody>
</table>

Figure 3.2: Relative approximation error of the estimated polynomial models (3.29a) [red] and (3.29b) [black] in the interval [−1, 6].

4. Although the condition number of the SVD_1 and SVD_{sc,1} solutions is quite large [of the order of 10^{16} > 10^{15}], the corresponding least squares predictions \( \hat{y}_{LS} \) are still accurate because of the numerical stable implementations (3.21b) and (3.28b).

Finally, the polynomial models (3.29), estimated via (3.28a), are evaluated over the interval [−1, 6] and compared with the arctan function. Figure 3.2 shows the corresponding relative approximation errors (3.32). It can be seen that the relative approximation errors become very large in the intervals [−1, 0) and (5, 6], which illustrates that an identified model should never be extrapolated. In other words, it should never be used outside the range of input values used for estimation.

Summarized, this polynomial curve fitting example has illustrated the importance of (i) the numerical solution method and (ii) the choice of the basis functions. A good indicator for the numerical stability of the solution method chosen, is the condition number of the regression matrix. Further, an identified model should never be extrapolated.
3.2.4. Weighted Linear Least Squares

A symmetric positive definite weighting \( W > 0 \), \( W \in \mathbb{R}^{N \times N} \) can be added to the least squares cost function (3.2)

\[
V_{WLS}(\theta, y) = \frac{1}{2} (y - H\theta)^T W (y - H\theta)
\] (3.33)

Minimization of (3.33) w.r.t. \( \theta \), gives

\[
\hat{\theta}_{WLS}(y) = (H^T WH)^{-1} H^T Wy
\] (3.34)

It can easily be verified that

\[
\mathbb{E}\{\hat{\theta}_{WLS}(y)\} = \theta_0
\] (3.35)

\[
\text{Cov}(\hat{\theta}_{WLS}(y)) = (H^T WH)^{-1} (H^T WC_v WH)(H^T WH)^{-1}
\] (3.36)

Hence, the weighted linear least squares estimator (3.34) is unbiased for any \( W \), and is efficient for normally distributed noise \( v \) if \( W = \lambda C_v^{-1} \), where \( \lambda > 0 \) [compare (3.36), with \( W = \lambda C_v^{-1} \), to (3.8)]. For any other choice of the weighting \( W \), the weighted linear least squares estimator is inefficient [proof: see Appendix 3.8 on page 61].

A numerical stable implementation of \( \hat{\theta}_{WLS}(y) \) (3.34), and the corresponding \( \hat{y}_{WLS} = H\hat{\theta}_{WLS}(y) \), consists in calculating the SVD of the weighted regression matrix \( W^{1/2}H = U_W \Sigma_W V_W^T \), where \( W^{1/2} \) is a (symmetric) square root of \( W \) [\( W^{T/2}W^{1/2} = W \)]. We find

\[
\hat{\theta}_{WLS}(y) = V_W \Sigma_W^{-1} U_W^T W^{1/2} y
\] (3.37a)

\[
\hat{y}_{WLS} = W^{-1/2} U_W U_W^T W^{1/2} y
\] (3.37b)

where (3.37b) uses \( \hat{y}_{WLS} = W^{-1/2}(W^{1/2}H)\hat{\theta}_{WLS}(y) \). A square root of a positive definite matrix can be calculated via a Cholesky factorization or the singular value decomposition (Golub and Van Loan, 1996).

3.3. The Linear Least Squares Estimator – Noisy Regression Matrix

The unbiasedness and consistency properties of the linear least squares estimator (3.3) have been proven assuming that the regression matrix \( H \) is known exactly and independent of the disturbing noise \( v \). In some applications this condition is not fulfilled. Two cases can be distinguished:
• **Case 1:** Only a noisy observation of the regression matrix $H$ is available

$$H = H_0 + \Delta H \quad \text{with} \quad E\{\Delta H\} = 0 \quad (3.38)$$

with $H_0$ the unknown noiseless part of $H$. In that case (3.1) becomes

$$y = H_0 \theta_0 + v \quad (3.39)$$

The second order moments are denoted as

$$E\{\Delta H^T v\} = C_{Hv} \quad (3.40a)$$

$$E\{\Delta H^T \Delta H\} = C_H \neq 0 \quad (3.40b)$$

where the cross-covariance $C_{Hv}$ can be zero.

• **Case 2:** The regression matrix $H$ is known exactly but it is correlated with the disturbing noise:

$$y = H \theta_0 + v \quad \text{with} \quad C_{Hv} = E\{H^T v\} \neq 0 \quad (3.41)$$

In the remainder of this section we handle both cases [Sections 3.3.1 and 3.3.2].

Next, we present three techniques for removing the asymptotic ($N \to \infty$) bias of the linear least squares method due to the noisy observation of the regression matrix $C_H \neq 0$ (3.40b) and/or the correlation between the regression matrix and the disturbing noise $C_{Hv} \neq 0$ (3.40a) and (3.41) [see Sections 3.3.3–3.3.5].

### 3.3.1. Case 1: Noisy Observation of the Regression Matrix

Due to the presence of $\Delta H$, the linear least squares estimate (3.3) is no longer unbiased and consistent. Indeed, combining (3.3a) with (3.39), we get

$$\hat{\theta}_{LS}(y) = (H^T H)^{-1}(H^T H_0)\theta_0 + (H^T H)^{-1}(H^T v) \quad (3.42)$$

where the first term in the right hand side of (3.42) is different from $\theta_0$, and where the expected value of the second term is not necessarily zero.

Under some assumptions the limit value of (3.42) can be calculated. If in addition to conditions 1–3 of Section 3.2.1, $\Delta H$ is mixing of order 4, then, $\hat{\theta}_{LS}(y)$ (3.42) converges in probability to the limit value

$$\text{plim}_{N \to \infty} \hat{\theta}_{LS}(y) = \lim_{N \to \infty} (H_0^T H_0 + C_H)^{-1}((H_0^T H_0)\theta_0 + C_{Hv}) \quad (3.43)$$

[proof: see Appendix 3.9 on page 61].
An example of a linear least squares problem with noisy observation of the regression matrix is the estimation of the resistor value from noisy DC current and voltage measurements (1.4b). It follows that (1.4a) is of the form (3.2), where

\[ y = [u(1), u(2), \ldots, u(N)]^T \] (3.44a)
\[ H = [i(1), i(2), \ldots, i(N)]^T \] (3.44b)
\[ \theta = R \] (3.44c)
\[ v = [n_u(1), n_u(2), \ldots, n_u(N)]^T \] (3.44d)

and with \( v \) the noise on \( y \). It can easily be verified that

\[ C_H v = 0 \] (3.45a)
\[ C_H = N\sigma_i^2 \] (3.45b)

where (3.45a) uses the independence of \( n_i(k) \) and \( n_u(k) \). Hence, the linear least squares estimator (1.4b) is not consistent, and its limit value (3.43) equals

\[ \lim_{N \to \infty} \left( \frac{Ni_0^2}{Ni_0^2 + N\sigma_i^2} R_0 \right) = \frac{R_0}{1 + \sigma_i^2/i_0^2} \]

which is consistent with (2.18).

### 3.3.2. Case 2: The Known Regression Matrix is Correlated with the Disturbing Noise

Combining (3.3a) with (3.41), we find

\[ \hat{\theta}_{LS}(y) = \theta_0 + (H^T H)^{-1}(H^T v) \] (3.46)

Following the same lines of Appendix 3.9 on page 61, we obtain the limit value of the (3.46)

\[ \operatorname{plim}_{N \to \infty} \hat{\theta}_{LS}(y) = \theta_0 + \lim_{N \to \infty} \left( \operatorname{E}(H^T H) \right)^{-1} C_{Hv} \] (3.47)

It follows that the linear least squares estimate (3.3) is inconsistent.

### 3.3.3. Bias Compensated Least Squares

If the second order moments (3.40) are known, then a bias compensated least squares (BCLS) can be constructed as

Case 1: \[ \hat{\theta}_{BCLS}(y) = \left( H^T H - C_H \right)^{-1}(H^T y - C_{Hv}) \] (3.48a)

Case 2: \[ \hat{\theta}_{BCLS}(y) = (H^T H)^{-1}(H^T y - C_{Hv}) \] (3.48b)
Figure 3.3: Comparison of the bias compensated least squares (3.50) [magenta] and the maximum likelihood solution (1.5b) [black]. For \( N = 2 \) the magenta line reaches 6Ω.

It can easily be verified that the bias compensated least squares estimator (3.48) is weakly consistent

\[
\lim_{N \to \infty} \hat{\theta}_{\text{BCLS}}(y) = \theta_0
\]

(proof: follow the same lines of Appendix 3.9 on page 61).

Applying (3.48a) to the estimation of the resistor value from noisy DC current and voltage measurements (3.44), we get

\[
\hat{R}_{\text{BCLS}}(N) = \frac{\sum_{k=1}^{N} i(k)u(k)}{\sum_{k=1}^{N} i^2(k) - N\sigma_i^2} = \frac{1}{N} \frac{\sum_{k=1}^{N} u(k)i(k)}{\sum_{k=1}^{N} i^2(k) - \sigma_i^2}
\]

(3.50)

which is consistent with (2.20). Figure 3.3 compares the BCLS estimate (3.50) to the ML estimate (1.5b). It can be seen that for large values of \( N \), the BCLS and ML estimates are very close to each other and converge to the true value 1Ω, while a significant difference can be observed for small values of \( N \).

The BCLS solution (3.48) has two drawbacks: (i) it requires the knowledge of the second order moments of the disturbing noise, and (ii) one must form the matrix product \( H^T H \), which might be ill-conditioned.

3.3.4. (Generalized) Total Least Squares

The total least squares method handles the first case [eqs. (3.38)–(3.40)] where a noisy observation \( H \) of the true regression matrix \( H_0 \) is available. The equa-
tion \( y \approx H\theta_0 \) is then rewritten as

\[
\begin{bmatrix}
    H & -y \\
    \tilde{H}
\end{bmatrix}
\begin{bmatrix}
    \theta_0 \\
    1
\end{bmatrix}
\approx 0
\]  

(3.51)

where \( \tilde{H} \in \mathbb{R}^{N \times (n_\theta+1)} \) is the augmented regression matrix and \( \tilde{\theta}_0 \in \mathbb{R}^{(n_\theta+1) \times 1} \) the augmented parameter vector, and with

\[
E\{\tilde{H}\} = \tilde{H}_0 = \begin{bmatrix} H_0 & -H_0\theta_0 \end{bmatrix} \Rightarrow \tilde{H}_0\tilde{\theta}_0 = 0
\]  

(3.52)

The total least squares solution \( \hat{\tilde{\theta}} \) of (3.51) is obtained as the right singular vector corresponding to the smallest singular value of the augmented regression matrix \( \tilde{H} \) (Van Huffel and Vandewalle, 1991; Markovsky and Van Huffel, 2007). Assuming that the singular values are ordered from large to small, we get

\[
\tilde{H} = U\Sigma V^T \Rightarrow \hat{\tilde{\theta}} = \tilde{V}_{[1:n_\theta]} \Rightarrow \hat{\theta}_{\text{TLS}}(y) = \hat{\theta}_{[1:n_\theta]} / \hat{\theta}_{[n_\theta+1]}
\]  

(3.53)

From the right hand side of (3.52), it follows that this solution gives the true model parameters \( \theta_0 \) when applied to the noiseless augmented regression matrix \( \tilde{H}_0 \). Assuming that the column covariance \( C_{\tilde{H}} = E\{\Delta\tilde{H}^T\Delta\tilde{H}\} \) of the augmented regression matrix \( \tilde{H} \) is proportional to the identity matrix

\[
C_{\tilde{H}} = \lambda I_{n_\theta+1}
\]  

(3.54)

the total least squares (TLS) estimate (3.53) is weakly consistent

\[
\underset{N \to \infty}{\text{plim}} \hat{\theta}_{\text{TLS}}(y) = \theta_0
\]  

(3.55)

[proof: see Appendix 3.10 on page 61].

If condition (3.54) is not satisfied, then the TLS estimate (3.53) is biased. A consistent estimate can be obtained as follows. First, a (symmetric) square root \( C_{\tilde{H}}^{1/2} \) of the column covariance \( C_{\tilde{H}} \) is calculated via a Cholesky or a singular value decomposition. For example,

\[
C_{\tilde{H}} = V_C\Sigma_C V_C^T \Rightarrow C_{\tilde{H}}^{1/2} = V_C\Sigma_C^{1/2} V_C^T
\]  

(3.56)

such that \( C_{\tilde{H}} = C_{\tilde{H}}^{T/2} C_{\tilde{H}}^{1/2} \). Next, (3.51) is rewritten as

\[
\tilde{H} C_{\tilde{H}}^{-1/2} C_{\tilde{H}}^{1/2} \tilde{\theta} \approx 0 \Rightarrow \tilde{H} C_{\tilde{H}}^{-1/2} \psi \approx 0
\]  

(3.57)
Finally, following the same lines of (3.53), the TLS estimate is obtained via the SVD of $\tilde{H}C_{\tilde{H}}^{-1/2}$ (Van Huffel and Vandewalle, 1991; Pintelon et al., 1998)

$$\tilde{H}C_{\tilde{H}}^{-1/2} = \tilde{U}\Sigma\tilde{V}^T \Rightarrow \hat{\psi} = \tilde{V}_{[:n_\theta+1]} \Rightarrow \hat{\theta}_{\text{GTLS}} = C_{\tilde{H}}^{-1/2}\hat{\psi}_{[1:n_\theta]} / \hat{\psi}_{[n_\theta+1]}$$

(3.58)

Solution (3.58) is called the generalized total least squares (GTLS) estimate. It is consistent because the column covariance of $\tilde{H}C_{\tilde{H}}^{-1/2}$ is equal to the identity matrix [proof: see Appendix 3.11 on page 62].

For some problems the column covariance of the augmented regression matrix is not of full rank. In that case $\tilde{H}C_{\tilde{H}}^{-1/2}$ in (3.57) is ill-defined. The GTLS estimate is then obtained via the generalized singular value decomposition (GSVD) of the matrix pair ($\tilde{H}$, $C_{\tilde{H}}^{1/2}$) (Pintelon et al., 1998; Pintelon and Schoukens, 2012)

$$\tilde{H} = U_1\Sigma_1X^{-1}$$

(3.59a)

$$C_{\tilde{H}}^{1/2} = U_2\Sigma_2X^{-1}$$

(3.59b)

where $\Sigma_i \in \mathbb{R}^{(n_\theta+1)\times(n_\theta+1)}$, $i = 1, 2$, are diagonal matrices with positive diagonal elements, $U_1 \in \mathbb{R}^{N\times(n_\theta+1)}$ and $U_2 \in \mathbb{R}^{(n_\theta+1)\times(n_\theta+1)}$ are orthogonal matrices, and with $X \in \mathbb{R}^{(n_\theta+1)\times(n_\theta+1)}$ (Golub and Van Loan, 1996). The ratios $\Sigma_{1[l,l]} / \Sigma_{2[l,l]}$, $l = 1, 2, \ldots, n_\theta + 1$, are called the generalized singular values. Using (3.59), the GTLS estimated is calculated as

$$\hat{\psi} = X_{[:k]} \Rightarrow \hat{\theta}_{\text{GTLS}} = \hat{\psi}_{[1:n_\theta]} / \hat{\psi}_{[n_\theta+1]}$$

(3.60)

where $k$ is the index corresponding to the smallest generalized singular value

$$k = \arg\min_l \frac{\Sigma_{1[l,l]}}{\Sigma_{2[l,l]}}$$

*Important note:* Matlab calculates the GSVD (3.59), where $X^{-1}$ is replaced by $X^T$. Hence, the $X$ of Matlab must first be transposed and inverted before applying (3.60).

The performance of the GTLS estimator (3.60) is illustrated on the estimation of the resistor value from noisy DC current and voltage measurements (3.44). The augmented regression matrix and the square root of its column covariance equal

$$\tilde{H} = \begin{bmatrix} i & -u \end{bmatrix}$$

(3.61)

$$C_{\tilde{H}}^{1/2} = \sqrt{N} \begin{bmatrix} \sigma_i & 0 \\ 0 & \sigma_u \end{bmatrix}$$

(3.62)
Figure 3.4: Comparison of the GTLS estimate (3.60) of the resistor value [green] with the bias compensated least squares (3.50) [magenta] and the maximum likelihood solution (1.5b) [black]. For \( N = 2 \) the magenta line reaches 6 \( \Omega \). For \( N = 1 \) the GTLS estimate [green] does not exist.

The GTLS estimates (3.60) for \( N > 1 \) are shown in Figure 3.4 [explain why it does not exist for \( N = 1 \)]. It can be seen that it performs better than the bias compensated least squares, and that it almost coincides with the maximum likelihood solution.

### 3.3.5. Instrumental Variables

A disadvantage of the bias compensated least squares (3.48) and generalized total least squares (3.60) is that the second order moments of the noise should be known. This problem is circumvented by the instrumental variables approach explained in the remainder of this section.

The asymptotic bias of the least squares solution (3.3a) originates from \( C_H \neq 0 \) (3.40b) and/or \( C_{Hv} \neq 0 \) (3.40a) and (3.41) [see (3.43) and (3.47)]. Assume now that we can construct a matrix \( G \) of the same size as \( H \) that is uncorrelated with the disturbing noise

\[
\mathbb{E}\{G^Tv\} = 0 \quad (3.63a)
\]
\[
\mathbb{E}\{G^T\Delta H\} = 0 \quad (3.63b)
\]

and correlated with the noiseless part of \( H \)

\[
\lim_{N \to \infty} \frac{1}{N} \mathbb{E}\{G^TH\} \text{ is of full rank} \quad (3.63c)
\]
Replacing $H^T$ in (3.3a) by $G^T$ defines the instrumental variables (IV) estimator
\[ \hat{\theta}_{IV}(y) = (G^T H)^{-1} G^T y \] (3.64)
(Norton, 1986). The signals used to construct the entries of the matrix $G$ are called the instrumental variables. Note that condition (3.63b) is not needed for Case 2 (3.41). Similar to the bias compensated least squares (3.48), the instrumental variables estimator (3.64) has the drawback that the matrix product $G^T H$ must be formed, which might be ill-conditioned.

Combining (3.64) with (3.39) [Case 1] and (3.64) with (3.41) [Case 2], gives

**Case 1** : \[ \hat{\theta}_{IV}(y) = (G^T H_0)^{-1} (G^T H_0)\theta_0 + (G^T H_0)^{-1} (G^T v) \] (3.65a)

**Case 2** : \[ \hat{\theta}_{IV}(y) = \theta_0 + (G^T H)^{-1} (G^T v) \] (3.65b)

Under conditions (3.63), (3.65) converges in probability to the true value
\[ \operatorname{plim}_{N \to \infty} \hat{\theta}_{IV}(y) = \theta_0 \] (3.66)
and, hence, (3.64) is weakly consistent [proof: follow the same lines of Appendix 3.9 on page 61]. For Case 2, some other classes of instrumental variable estimators can be found in Söderström (2013).

The basic idea of the instrumental variables method is now illustrated on the estimation of the resistor value from noisy DC current and voltage measurements (3.44). In a first attempt, we use the first $N-1$ current and voltage samples as data for $y$ and $H$, and the last $N-1$ current samples as instrumental variables for constructing $G$

\[ y = [u(1), u(2), \ldots, u(N-1)]^T \] (3.67a)
\[ H = [i(1), i(2), \ldots, i(N-1)]^T \] (3.67b)
\[ G = [i(2), i(3), \ldots, i(N)]^T \] (3.67c)

Since the current measurements are independently distributed and independent of the voltage measurements, $G$ (3.67c) satisfies conditions (3.63), and the resulting IV estimate (3.64) equals
\[ \hat{R}_{IV}(N) = \frac{\sum_{k=1}^{N-1} u(k)i(k+1)}{\sum_{k=1}^{N-1} i(k)i(k+1)} = \frac{1}{N-1} \sum_{k=1}^{N-1} u(k)i(k+1) \] (3.68)

[Exercise: using the strong law of large numbers (2.8a), prove that the IV estimate (3.68) is strongly consistent]. A drawback of (3.68) is that not all voltage measurements are used.
Therefore, in a second attempt, we split the current and voltage data records in two equal parts, giving

\[ y = \begin{bmatrix} y_1 \\ y_2 \end{bmatrix} \quad \text{and} \quad H = \begin{bmatrix} H_1 \\ H_2 \end{bmatrix} \]  

(3.69a)

\[ y_1 = [u(1), u(2), \ldots, u(N/2)]^T \]  

(3.69b)

\[ y_2 = [u(N/2 + 1), u(N/2 + 2), \ldots, u(N)]^T \]  

(3.69c)

\[ H_1 = [i(1), i(2), \ldots, i(N/2)]^T \]  

(3.69d)

\[ H_2 = [i(N/2 + 1), i(N/2 + 2), \ldots, i(N)]^T \]  

(3.69e)

Using the current and voltage signals of the second part as instrumental variables for the first part and vice versa, we get for \( G \)

\[ G = \begin{bmatrix} H_2 \\ H_1 \end{bmatrix} \]  

(3.70)

It can easily be verified that \( G \) (3.70) satisfies conditions (3.63), and the resulting IV estimate (3.64) equals

\[
\hat{R}_{IV}(N) = \frac{\sum_{k=1}^{N/2} u(k)i(k+N/2) + u(k+N/2)i(k)}{2 \sum_{k=1}^{N/2} i(k)i(k+N/2)} \\
= \frac{2}{N} \sum_{k=1}^{N/2} u(k)i(k+N/2) + u(k+N/2)i(k)
\]

(3.71)

[Exercise: using the strong law of large numbers (2.8a), prove that the IV estimate (3.71) is strongly consistent]. Using the first order Taylor series method of Section 2.3.2 on page 22, the asymptotic variance of (3.71) is found to be

\[
\text{var}(\hat{R}_{IV}(N)) = \frac{R_0^2}{N} \left( \frac{\sigma_u^2}{u_0^2} + \frac{\sigma_i^2}{i_0^2} + \frac{\sigma_u^2 \sigma_i^2}{u_0^2 i_0^2} + 2 \frac{\sigma_i^4}{i_0^4} \right)
\]

(3.72)

which is exactly the same as that of the bias compensated least squares estimator (2.36) [Exercise: prove (3.72)].

Figure 3.5 compares the IV estimate (3.71), with the bias compensated least squares (3.50) and the maximum likelihood solution (1.5b). In contrast to \( \hat{R}_{BCLS}(N) \), \( \hat{R}_{IV}(N) \) is everywhere close to \( \hat{R}_{ML}(N) \). For large values of \( N \), \( \hat{R}_{BCLS}(N) \) and \( \hat{R}_{IV}(N) \) perform equally well, while their small sample behavior is quite different.

From the resistor example and the general theory it is clear that the choice of the instrumental variables is not unique, and might strongly impact the
covariance of the estimates. Similar to the weighted linear least squares (3.34), a positive definite weighting $W$ can be added to the instrumental variables estimate (3.64):

$$\hat{\theta}_{WIV}(y) = (G^TWH)^{-1}G^Ty$$

(3.73)

How to choose the weighting $W$ to minimize the covariance is discussed in Söderström and Stoica (1989) for the Case 2 least squares problem [see Section 3.3.2].

### 3.4. Regularized Linear Least Squares

In Section 1.2.2 on page 6 it has been shown that increasing the number of model parameters decreases the bias but increases the variability of the estimated model [see Figure 1.5 on page 7]. In this section we present a method that makes automatically a bias–variance trade off. Compared with the classical AIC and MDL model selection criteria [see Section 8.3 on page 132], (i) the estimation of the model parameters and the tuning of the model complexity are performed in one step, and (ii) the complexity is tuned in a continuous manner.

The basic idea of regularization consists in adding a penalty term for model complexity to the linear least squares cost function (3.2). Different choices for the penalty (regularization) term are possible. If a quadratic form in $\theta$ is chosen,
then this results in the regularized linear least squares method
\[
V_{\text{RLS}}(\theta, y) = \frac{1}{2} (y - H\theta)^T (y - H\theta) + \frac{1}{2} \gamma \theta^T P^{-1} \theta \tag{3.74}
\]
where \( P \) is a symmetric positive (semi-)definite matrix – called the regularization matrix – and \( \gamma \geq 0 \). If \( P \) is not of full rank, then \( P^{-1} \) in (3.74) is replaced by the Moore-Penrose pseudo-inverse \( P^+ \) (Ben-Israel and Greville, 1974).

Minimizing (3.74) w.r.t. \( \theta \), gives the regularized linear least squares (RLS) estimate
\[
\hat{\theta}_{\text{RLS}}(y) = (PH^T H + \gamma I_{n_\theta})^{-1} P H^T y \tag{3.75}
\]
Note that (3.75) is well defined for singular regularization matrices \( P \), and it is still the optimal solution (Pillonetto et al., 2014). Following the same lines of (3.25), a numerical stable implementation of (3.75) is,
\[
\hat{\psi}_{\text{RLS}}(y) = (PH^T H + \gamma I_{n_\theta})^{-1} P H^T y \quad \text{(3.76a)}
\]
\[
\hat{\theta}_{\text{RLS}}(y) = S^{-1} \hat{\psi}_{\text{RLS}}(y) \quad \text{(3.76b)}
\]
where \( S \) is defined in (3.25), and where the second term in (3.74) is replaced by \( 0.5 \gamma \psi^T P^{-1} \psi \).

Assuming that the disturbing noise is white, \( \text{Cov}(v) = \sigma_v^2 I_N \), the mean squared error (MSE) of (3.75) equals
\[
\text{MSE}(\hat{\theta}_{\text{RLS}}) = (PH^T H + \gamma I_{n_\theta})^{-1} (\sigma_v^2 P H^T H P + \gamma^2 \theta_0 \theta_0^T)(H^T H P + \gamma I_{n_\theta})^{-1} \tag{3.77}
\]
It can be shown that the mean squared error (3.77) is minimal for the choice
\[
\gamma = \sigma_v^2 \quad \text{(3.78a)}
\]
\[
P = \theta_0 \theta_0^T \quad \text{(3.78b)}
\]
in (3.74) [proof: see Chen et al., 2012, Section 6]. Note that the optimal \( P \) matrix has rank one and depends on the true unknown model parameters \( \theta_0 \).

Substituting in (3.74)
\[
\gamma = \sigma_v^2 \tag{3.79a}
\]
\[
P = I_{n_\theta} \tag{3.79b}
\]
one obtains the \( L_2 \) or Tikhonov regularization (Tikhonov and Arsenin, 1977). Note that the choice of \( \gamma \) (3.79a) minimizes the mean squared error (3.77) where \( P \) is replaced by (3.79b) [Exercise: prove this statement. Hint: calculate
the derivative of (3.77) w.r.t. $\gamma$ and show that $\gamma = \sigma^2_v$ is a solution of the equation. Note also that the choice (3.79b) implicitly assumes that all entries $\theta_{[i]}$ have the same physical units, which – in general – is not true. Therefore, the columns of the regression matrix $H$ must first be scaled by their 2-norm as in (3.25) such that the unknown model parameters have no units.

The performance of the regularized linear least squares estimator (3.75) is illustrated on the polynomial curve fitting example of Section 1.2.2 on page 6. We estimate a polynomial $y(u, \theta)$ of order $n_\theta - 1 = 14$

$$y(u, \theta) = \sum_{r=1}^{n_\theta} \theta_{[r]} (u - 1)^{r-1}$$  \hspace{1cm} (3.80)

from $N = 17$ noisy samples $y(k)$, $k = 1, 2, \ldots, N$, of the arctan function (1.6) over the interval $[-2, 4]$

$$u(k) = \frac{6}{N - 1} (k - 1) - 2$$  \hspace{1cm} (3.81)

Since the true parameters values $\theta_0$ are unknown, we first estimate (3.80) from the noiseless arctan values using the SVD solution (3.28) of the scaled linear least squares problem (3.25) where

$$H_{[k,r]} = (u(k) - 1)^{r-1}$$  \hspace{1cm} (3.82)

and with $k = 1, 2, \ldots, N$ and $r = 1, 2, \ldots, n_\theta$. The left column of Figure 3.6 shows the approximation error. Next, the regularized linear least squares estimate (3.76) is calculated from the noisy samples for the following three choices:

1. No regularization: $\gamma = 0$.
2. Tikhonov regularization: $P = I_{n_\theta}$ and $\gamma = \sigma^2_y$.
3. Optimal regularization: $P = \theta_0 \theta_0^T$ and $\gamma = \sigma^2_y$.

This is repeated for $M = 1000$ independent disturbing noise realizations $n_y(k)$, $k = 1, 2, \ldots, N$, in (1.6). Figure 3.6 shows the sample estimate (1.8) of the mean squared error [bottom right plot] over the $M = 1000$ Monte-Carlo runs, and the estimates (3.76) for the last Monte-Carlo run. It can be seen that:

1. The Tikhonov (3.79) and optimal (3.78) regularizations perform better than the ordinary linear least squares.
2. The optimally regularized least squares solution with $n_\theta - 1 = 14$ performs better than the lower order model $n_\theta - 1 = 3$ in Figure 1.5 on page 7. This can be explained by the fact that the regularized linear least squares method (3.75) tunes in a continuous way the model complexity.
Figure 3.6: Regularized polynomial fitting (3.75) of order \( n_\theta - 1 = 14 \) from \( N = 17 \) noisy samples (black \( \times \)) of the arctan function for different choices of the matrix \( P \) and the scalar \( \gamma \). Yellow: linear least squares, \( P = I_{n_\theta} \) and \( \gamma = 0 \). Red: Tikhonov regularized linear least squares, \( P = I_{n_\theta} \) and \( \gamma = \sigma_y^2 \). Blue: optimally regularized linear least squares, \( P = \theta_0\theta_0^T \) and \( \gamma = \sigma_y^2 \). Left column: \( \sigma_y = 0 \). Right column: \( \sigma_y = 0.5 \). Top row: results of the last Monte-Carlo run. Bottom left: bias error \( \hat{y}(k) - \arctan(u(k)) \). Bottom right: root mean squared error [square root of (1.8)] of the \( M = 1000 \) Monte-Carlo runs.

A drawback of the optimally regularized linear least squares method is the need for the prior knowledge \( \sigma_v^2 \) and \( \theta_0 \), which is mostly not available. How to handle this important issue is explained in Section 6.4, where the connection between (3.75) and Gaussian process regression is established. In Gaussian process regression the matrix \( P \), called kernel matrix, is parametrized, and these parameters, together with \( \gamma \), are tuned such that an optimal trade-off between data-fit and model complexity is made.

### 3.5. Outliers

Due to sensor malfunctioning and/or data transmission errors outliers can be present in measured data. The best thing to do is to solve the hardware problem and to redo the experiments. Sometimes this is not possible because of time and cost, or because the experiments cannot be redone, e.g. in case of
historical data.

The (weighted) least squares cost functions (3.2) and (3.33) emphasize outliers, resulting in an increased variability of the (weighted) linear least squares estimates (3.3) and (3.34). A first solution consists in detecting the outliers and removing them from the data. If this is not possible, then, the only option left is to make the cost function less sensitive to outliers. Among the different robustifications of the least squares cost function, Huber’s robust norm (Huber, 1981) is popular

\[ V_{\text{Huber}}(\theta, y) = \sum_{k=1}^{N} L(e[k])(\theta, y) \]  

where \( \delta \) is the absolute value of the residual \( e \) above which the quadratic function is replaced by the absolute value [see Figure 3.7]. Proceeding in this way, outliers contribute less to the cost function.

An appropriate value for \( \delta \) in (3.83) is obtained via a two-step procedure. First, the linear least squares estimate (3.3) is calculated using one of the numerical stable implementations. Next, a robust estimate \( \hat{\sigma}_e \) of the standard deviation of the linear least squares residuals \( e[k](\hat{\theta}_{\text{LS}}(y), y) \), \( k = 1, 2, \ldots, N \),
where $e_{k}(\theta, y)$ is defined in (3.83c), is given by

$$
\hat{\sigma}_{e} = \frac{1}{0.6745} \text{median}_{k} \left( |e_{\text{LS},[k]} - \text{median}_{k}(e_{\text{LS},[k]}) | \right)
$$

(3.84a)

$$
e_{\text{LS},[k]} = e_{[k]}(\hat{\theta}_{\text{LS}}(y), y)
$$

(3.84b)

(Hampel et al., 2005). Finally, $\delta$ is chosen as

$$
\delta = \rho \hat{\sigma}_{e} \quad \text{with } \rho \in [1, 1.8]
$$

(3.85)

and the Huber cost function (3.83) is minimized using convex optimization techniques (Boyd and Vandenberghe, 2004).

As an example we retake the polynomial approximation of the arctan function (3.80)–(3.82), where $N = 100$ and $n_{\theta} - 1 = 7$. First, zero mean Gaussian noise with standard deviation $\sigma_{y} = 0.2$ is added to the true arctan function values, and the linear least squares (3.28) and robust Huber [minimizer (3.83)] estimates are calculated. The latter is done by applying the Matlab function “robustfit” on the scaled regression matrix $HS^{-1}$ [see (3.25)]. As default value, $\rho = 1.345$ in (3.85) is used in the Matlab function. This is repeated for $M = 400$ independent disturbing noise realizations. Figure 3.8, first two rows, shows the results. As expected, the linear least squares and robust Huber estimates perform equally well.

Next, for each of the $M = 400$ noise realizations, randomly distributed outliers are added, with amplitude 5 and equally distributed random sign. The fraction of the outliers equals 30% [0.3N]. Figure 3.8, last two rows, shows the results. A significant difference in performance between the linear least squares and robust Huber estimates can now be observed. This is due to the high sensitivity of the linear least squares cost function (3.2) to outliers.

Finally, the linear least squares estimator is applied to each of the $M = 400$ data sets where the samples with the outliers have been removed. The resulting RMS error [green line in the bottom left plot of Figure 3.8] is much lower than that of the linear least squares and robust Huber estimates over the full data sets with outliers, and only about 1.5 dB higher than that of the full data sets without outliers [second row of Figure 3.8]. The latter can be explained by the ratio of the number of samples $N$ of the full data set to the number of samples $0.7N$ of the data set with the outliers removed: $\text{db}(N/(0.7N))/2 = 1.55$.
Figure 3.8: Polynomial approximation of order \( n_\theta - 1 = 7 \) of the arctan function disturbed by zero mean white Gaussian noise with standard deviation \( \sigma_y = 0.2 \) [first two rows] plus 30\% fraction of randomly distributed outliers with random sign [last two rows]. Left: linear least squares estimate (3.28). Right: Robust Huber estimate [minimizer (3.83)]. First and third rows: true value [black], and disturbed data [blue] and estimate [red] of the last Monte-Carlo run. Second and fourth rows: RMS error of the estimates over the \( M = 1000 \) Monte-Carlo runs using all data [blue] or with outliers removed [green]. For the third row, only the \( y \)-values in the range \([-2, 2]\) are shown.
3.6. Appendix: Schwarz Inequality for Matrices

Consider the matrices $A \in \mathbb{R}^{N \times n_a}$ and $B \in \mathbb{R}^{N \times n_b}$, with $N > n_a, n_b$ and where $A$ has full rank $n_a$. We will show that

$$B^T B \geq (B^T A)(A^T A)^{-1}(A^T B)$$

(3.86)

which is the equivalent of (2.64) for matrices.

Proof: Consider the singular value decomposition

$$A = U \Sigma V^T$$

(3.87)

where $U \in \mathbb{R}^{N \times n_a}$ is a rectangular orthogonal matrix ($U^T U = I_{n_a}$, with $I_{n_a}$ the $n_a \times n_a$ identity matrix), $\Sigma$ a diagonal $n_a \times n_a$ matrix containing the singular values, and $V \in \mathbb{R}^{n_a \times n_a}$ a square orthogonal matrix ($V^{-1} = V^T$). Using (3.87), where $\Sigma$ is of full rank, the matrix product $A(A^T A)^{-1}A^T$ in (3.86) can be elaborated as

$$A(A^T A)^{-1}A^T = U \Sigma V^T$$

(3.88)

Defining $U^\perp \in \mathbb{R}^{N \times (N-n_a)}$ as the orthogonal complement of $U$, the $N \times N$ identity matrix $I_N$ can be written as

$$I_N = [UU^\perp][UU^\perp]^T = UU^T + U^\perp U^\perp^T$$

(3.89)

Using (3.88) and (3.89), inequality (3.86) can be rewritten as

$$B^T U U^T B + B^T U^\perp U^\perp^T B \geq B^T U U^T B \Leftrightarrow (U^\perp^T B)^T(U^\perp^T B) \geq 0$$

The last inequality in (3.90) is indeed valid, which proves the Schwarz inequality (3.86).

3.7. Appendix: The Linear Least Squares Estimator is Inefficient

We must show that

$$(H^T C_v^{-1}H)^{-1} \leq (H^T H)^{-1}(H^T C_v H)(H^T H)^{-1}$$

(3.90)

or, equivalently, by taking the inverse of both sides

$$(H^T C_v^{-1}H) \geq (H^T H)(H^T C_v H)^{-1}(H^T H)$$

(3.91)

Using the Cholesky decomposition $C_v = C^T C$, and making the choice $B = C^{-T}H$ and $A = CH$ in (3.86), proves (3.91) and, hence, also the inefficiency of the linear least squares method.
3.8. Appendix: The Weighted Linear Least Squares Estimator is Inefficient for any Choice $W \neq \lambda C_v^{-1}$

We must demonstrate that
\[(H^T C_v^{-1} H)^{-1} \leq (H^T WH)^{-1}(H^T W C_v W H)(H^T WH)^{-1} \quad (3.92)\]
for any $W \neq \lambda C_v^{-1}$, with $\lambda > 0$, or, equivalently, by taking the inverse of both sides
\[(H^T C_v^{-1} H) \geq (H^T WH)(H^T W C_v W H)^{-1}(H^T WH) \quad (3.93)\]
Using the Cholesky decomposition $C_v = C^T C$, and making the choice $B = C^{-T} H$ and $A = C W H$ in (3.86), proves (3.93) and, hence, also the inefficiency of the weighted linear least squares method. Only when $W = \lambda C_v^{-1}$, with $\lambda > 0$, the right and left hand sides of (3.92) are equal.

3.9. Appendix: Proof of the Limit Value (3.43)

Since $\Delta H$ has zero mean value and is mixing of order 4, we can apply the law of large numbers (2.8c) to each of the sums in (3.42), giving
\[1.\text{i.m. } \frac{1}{N} \frac{1}{N}(H^T v - C_H v) = 0 \quad (3.94a)\]
\[1.\text{i.m. } \frac{1}{N} (H^T H - C_H) = 0 \quad (3.94b)\]
\[1.\text{i.m. } \frac{1}{N} (H^T H_0 - H_0^T H_0) = 0 \quad (3.94c)\]
where we used notations (3.40), $\mathbb{E}\{\Delta H\} = 0$, and the independence of $\Delta H$ and $H_0$. Since convergence in mean square sense (3.94) implies convergence in probability [see Figure 2.1 on page 14], and since the limit in probability and a continuous function can be interchanged, the limit value (3.43) follows immediately from (3.42) and (3.94).

3.10. Appendix: Weak Consistency of the Total Least Squares Estimator (3.55)

Since $\Delta \tilde{H}$ has zero mean value and is mixing of order 4, we can apply the law of large numbers (2.8c) to $\tilde{H}^T \tilde{H}$, giving
\[1.\text{i.m. } \frac{1}{N} (\tilde{H}^T \tilde{H} - \mathbb{E}\{\tilde{H}^T \tilde{H}\}) = 0 \quad (3.95)\]
The expected value in (3.95) is elaborated as

$$E\{\tilde{H}^T \tilde{H}\} = \tilde{H}_0^T \tilde{H}_0 + C_{\tilde{H}}$$  \hspace{1cm} (3.96)$$

where $C_{\tilde{H}} = E\{\Delta \tilde{H}^T \Delta \tilde{H}\}$ is the column covariance of $\tilde{H}$. Combining (3.96) with (3.54) and the singular value decomposition $\tilde{H}_0 = \tilde{U}_0 \tilde{\Sigma}_0 \tilde{V}_0^T$, we find

$$E\{\tilde{H}^T \tilde{H}\} = \tilde{V}_0 \tilde{\Sigma}_0^2 \tilde{V}_0^T + \lambda I_{n_\theta+1}$$
$$= \tilde{V}_0 (\tilde{\Sigma}_0^2 + \lambda I_{n_\theta+1}) \tilde{V}_0^T$$  \hspace{1cm} (3.97)$$

where the second equality uses $\tilde{V}_0 \tilde{V}_0^T = I_{n_\theta+1}$. Since the eigenvectors of $\tilde{H}^T \tilde{H}$ equal the right singular vectors of $\tilde{H}$, it follows from (3.95) and (3.97) that the right singular vectors of $\tilde{H}$ converge in probability to the right singular vectors of $\tilde{H}_0$

$$\text{plim}_{N \to \infty} \tilde{V} = \tilde{V}_0$$  \hspace{1cm} (3.98)$$

Combining (3.98) with (3.53), proves (3.55).

### 3.11. Appendix: Weak Consistency of the Generalized Total Least Squares Estimator (3.58)

The column covariance of $\tilde{H} C_{\tilde{H}}^{-1/2}$ can be elaborated as

$$E\{(\Delta \tilde{H} C_{\tilde{H}}^{-1/2})^T (\Delta \tilde{H} C_{\tilde{H}}^{-1/2})\} = C_{\tilde{H}}^{-T/2} E\{\Delta \tilde{H}^T \Delta \tilde{H}\} C_{\tilde{H}}^{-1/2}$$
$$= C_{\tilde{H}}^{-T/2} C_{\tilde{H}} C_{\tilde{H}}^{-1/2}$$
$$= C_{\tilde{H}}^{-T/2} C_{\tilde{H}}^{1/2} C_{\tilde{H}}^{-1/2}$$
$$= I_{n_\theta+1}$$

We conclude that condition (3.54) is fulfilled with $\lambda = 1$. Hence, the GTLS estimate (3.58) is weakly consistent.
Chapter 4
Nonlinear Least Squares

Abstract: The main difficulties of nonlinear regression problems are (i) the generation of initial estimates and (ii) the nonlinear minimization. This chapter studies the stochastic properties of nonlinear regression problems, gives an overview of the main nonlinear minimization algorithms, and presents some guidelines for generating starting values.

Learning Objectives:

- Derivation of the asymptotic properties of (weighted) nonlinear least squares estimators.
- Nonlinear minimization algorithms and their numerical stable implementations.
- Separable nonlinear least squares.
- Generation of starting values.

4.1. Problem Statement

In this chapter we consider nonlinear regression problems of the form

\[ y = H(\theta_0) + v \]  

with \( \theta_0 \in \mathbb{R}^{n_{\theta} \times 1} \) the true unknown model parameters, \( y \in \mathbb{R}^{N \times 1} \) the noisy data, \( H(\theta) \in \mathbb{R}^{N \times 1} \) a nonlinear vector function of \( \theta \), and where the disturbing noise \( v \in \mathbb{R}^{N \times 1} \) has zero mean value and covariance \( C_v \). We assume furthermore that the vector function \( H(\theta) \) is noiseless. Note that this condition is not always fulfilled [see Chapter 5 for a practical example].

The model parameters \( \theta \) are obtained by minimizing

\[ V_{\text{NLS}}(\theta, y) = \frac{1}{2} (y - H(\theta))^T (y - H(\theta)) \]  

(4.2)
w.r.t. \( \theta \), which is a nonlinear minimization problem. This issue is tackled in Sections 4.4 and 4.5. The minimizer of (4.2) is called the nonlinear least squares estimate \( \hat{\theta}_{\text{NLS}}(y) \). Following the same lines of Section 3.2.4 on page 44, a symmetric positive definite weighting matrix \( W \in \mathbb{R}^{N \times N} \) can be added to the cost function

\[
V_{\text{WNLS}}(\theta, y) = \frac{1}{2} (y - H(\theta))^T W (y - H(\theta))
\]

The goal is to choose \( W \) such that the asymptotic covariance of the weighted nonlinear least squares (WNLS) estimate \( \hat{\theta}_{\text{WNLS}}(y) \) is smaller than that of \( \hat{\theta}_{\text{NLS}}(y) \). This problem is handled in Section 4.2.3.

### 4.2. Stochastic Properties

Since no analytic expressions for \( \hat{\theta}_{\text{NLS}}(y) \) and \( \hat{\theta}_{\text{WNLS}}(y) \) are available as a function of \( y \), their asymptotic (\( N \to \infty \)) are analyzed following the lines of Section 2.4 on page 28. It requires the following technical conditions:

1. \( H(\theta) \) is a continuous function of \( \theta \) with continuous first and second order derivatives. This condition should be fulfilled in a closed and bounded neighborhood of \( \theta_0 \).

2. Adding data should add information about \( \theta \)

\[
\lim_{N \to \infty} \frac{1}{N} \left( \frac{\partial H(\theta)}{\partial \theta} \right)^T W \left( \frac{\partial H(\theta)}{\partial \theta} \right) \quad \text{is of full rank} \tag{4.4}
\]

in a closed and bounded neighborhood of \( \theta_0 \).

3. For any \( N \), infinity included, the weighting matrix \( W \) should have a finite 1-norm

\[
\max_i \sum_{k=1}^N |W_{[k,i]}| \leq c < \infty \tag{4.5}
\]

where the constant \( c \) is independent of \( N \) [see Pintelon and Schoukens, 2012, page 629]. The interpretation of this condition will become clear in Section 4.2.3.

4. The correlation of the noise over the measurements should decrease sufficiently fast to zero. This is mathematically translated as a mixing condition of a certain order \( P \geq 2 \). For example, it implies that the off-diagonal elements of \( C_v \) decrease sufficiently fast to zero as the distance to the main diagonal increases.
An example of a nonlinear least squares problem that does not satisfy condition (4.4) is
\[ y[k] = Ae^{-(k-1)\tau} \cos(\omega(k-1)T_s + \phi) + v[k] \] (4.6)
where \( \tau > 0 \) and \( k = 1, 2, \ldots, N \). The reason for this is that the signal part of (4.6) decays exponentially to zero as \( N \to \infty \). Hence, gathering more data (increasing \( N \)) will not give additional information about the signal parameters.

4.2.1. Consistency

Referring to Section 2.4.1 on page 29, the weighted nonlinear least squares estimator is weakly consistent if (i) the expected value of the cost function (4.3) is minimal in the true model parameters \( \theta_0 \), and (ii) the cost function converges uniformly in probability to its expected value in a closed and bounded neighborhood of \( \theta_0 \). Both conditions are checked in the remainder of this section.

Calculating the expected value of the weighted nonlinear least squares cost function (4.3), we find,
\[
V_{\text{WNLS}}(\theta) = \mathbb{E}\{V_{\text{WNLS}}(\theta, y)\} = \frac{1}{2} \mathbb{E}\left\{ (H(\theta_0) - H(\theta) + v)^T W (H(\theta_0) - H(\theta)) + \frac{1}{2} \text{trace}(Wv^T) \right\} \]
(4.7)
where the first equality uses (4.1), and the second equality
\[
v^T W v = \text{trace}(v^T W v) = \text{trace}(Wv^T) \]
with \( \text{trace}(A) = \sum_k A[k,k] \). It can be seen that the first term in (4.7) is zero in \( \theta = \theta_0 \) and that the second term is independent of \( \theta \). Hence, \( \mathbb{E}\{V_{\text{WNLS}}(\theta, y)\} \) is minimal in the true model parameters.

Under technical conditions 1, 3 and 4, with \( P = 4 \), the cost (4.3) converges uniformly in mean squared sense to its expected value [see Pintelon and Schoukens, 2012, Chapter 17]. This implies convergence in probability [see Figure 2.1 on page 14]. Together with (4.7) it establishes the weak consistency of weighted nonlinear least squares estimator
\[
\lim_{N \to \infty} \hat{\theta}_{\text{WNLS}}(y) = \theta_0 \] (4.8)
[Question: Why do we need a mixing condition of order four instead of order two as for the linear least squares estimator?]
4.2.2. Asymptotic Covariance

Direct application of (2.55) and (2.56) to (4.3), gives

\[ \hat{\theta}_{WNLS}(y) - \theta_0 \xrightarrow{\text{in prob.}} \delta_{\theta,WNLS}(v) = -(H_0^T W H_0')^{-1} H_0^T W v \]  

(4.9a)

\[ H_0' = \left. \frac{\partial H(\theta)}{\partial \theta} \right|_{\theta=\theta_0} \quad \text{and} \quad H_0 = H(\theta_0) \]  

(4.9b)

and

\[ \text{Cov}(\hat{\theta}_{WNLS}(y)) = \text{Cov}(\delta_{\theta,WNLS}(v)) = (H_0^T W H_0')^{-1} H_0^T W C_v W H_0' (H_0^T W H_0')^{-1} \]  

(4.10)

[proof: see Appendix 4.6 on page 78].

4.2.3. Asymptotic Efficiency

Assuming that the noise \( v \) is normally distributed, the pdf of the measurements \( y \), given the true model parameters \( \theta_0 \), is given by

\[ f_{y|\theta_0}(y|\theta_0) = \frac{1}{\sqrt{(2\pi)^N \det C_v}} e^{-\frac{1}{2}(y-H(\theta_0))' C_v^{-1} (y-H(\theta_0))} \]  

(4.11)

Using (4.11), the Fisher information matrix \( F_i(2.16b) \) is readily found

\[ F_i(\theta_0) = H_0^T C_v^{-1} H_0' \]  

(4.12)

where \( H_0' \) is defined in (4.9b). This results in the following Cramér-Rao lower bound

\[ \text{Cov}(\hat{\theta}(y)) \geq (H_0^T C_v^{-1} H_0')^{-1} \]  

(4.13)

Comparing (4.10) and (4.13) it can be seen that the weighted nonlinear least squares estimator is asymptotically efficient for normally distributed noise \( v \) if \( W = \lambda C_v^{-1} \), where \( \lambda > 0 \). For any other choice of the weighting \( W \), the weighted nonlinear least squares estimator is inefficient [proof: follow the same lines of Appendix 3.8 on page 61, where \( H \) is replaced by \( H_0' \)].

4.2.4. Asymptotic Normality

If technical conditions 1–4 are fulfilled with \( P = \infty \), then, \( \delta_{\theta,WNLS}(v) \) in (4.9a) is also mixing of order infinity [see Pintelon and Schoukens, 2012, p. 575, Lemma 16.5]. Hence, the third version of the central limit theorem in Section 2.2.5 on page 16 can be applied to \( \delta_{\theta,WNLS}(v) \), which shows that \( \hat{\theta}_{WNLS}(y) \) is asymptotically normally distributed, with mean value \( \theta_0 \) and asymptotic covariance (4.10).
4.3. Separable Nonlinear Least Squares

Some nonlinear models are linear in a subset of the model parameters. These models can be written as

\[ y = H(\theta)C + v \]  \hspace{1cm} (4.14)

with \( C \in \mathbb{R}^{n_c \times 1} \), \( H(\theta) \in \mathbb{R}^{N \times n_c} \), \( y \in \mathbb{R}^{N \times 1} \), and \( v \in \mathbb{R}^{N \times 1} \) zero mean disturbing noise with covariance \( C_v \). The corresponding separable nonlinear least squares (SNLS) cost function equals

\[ V_{SNLS}(\theta, C, y) = \frac{1}{2} (y - H(\theta)C)^T(y - H(\theta)C) \]  \hspace{1cm} (4.15)

In Section 4.3.1 we show that \( C \) can analytically be eliminated in (4.15), which simplifies the nonlinear minimization problem (Golub and Pereyra, 1973). This procedure is called the variable projection method (Golub and Pereyra, 2003). Next, a practical example is given [Section 4.3.2]. Note that the variable projection method is also applicable if a symmetric positive definite weighting matrix \( W \) is added to the cost function (4.15).

4.3.1. Variable Projection Method

Minimization of (4.15) w.r.t. \( C \) gives

\[ C = (H^T(\theta)H(\theta))^{-1}H^T(\theta)y \]  \hspace{1cm} (4.16)

Substituting (4.16) in the cost function (4.15), we find

\[ V_{SNLS}(\theta, y) = \frac{1}{2} (P(\theta)y)^T(P(\theta)y) \]  \hspace{1cm} (4.17a)

\[ P(\theta) = I_N - H(\theta)(H^T(\theta)H(\theta))^{-1}H^T(\theta) \]  \hspace{1cm} (4.17b)

where \( P(\theta) \) is a symmetric idempotent projection matrix: \( P^2 = P \). This projection matrix is calculated in a numerical reliable way via the SVD of \( H(\theta) \)

\[ H(\theta) = U(\theta)\Sigma(\theta)V^T(\theta) \quad \text{and} \quad U^\perp(\theta) \]  \hspace{1cm} (4.18)

where \( U^\perp(\theta) \in \mathbb{R}^{N \times (N-n_c)} \) is the orthogonal complement of \( U(\theta) \in \mathbb{R}^{N \times n_c} \) [the full SVD in Matlab provides the orthogonal complement \( U^\perp(\theta) \)]. Using (4.18), the cost function (4.17a) and the projection matrix (4.17b) are calculated as

\[ V_{SNLS}(\theta, y) = \frac{1}{2} (U^\perp(\theta)y)^T(U^\perp(\theta)y) \]  \hspace{1cm} (4.19a)

\[ P(\theta) = U^\perp(\theta)U^\perp T(\theta) \]  \hspace{1cm} (4.19b)
[proof: use $I_N = [U^TU][U^TU]^T$ and $U^TU = I_{N-n_c}$]. Following the same lines, the derivative of $P(\theta)$ w.r.t. the model parameters $\theta$ can also be calculated in a numerical stable manner.

From (4.17a), it can be seen that the data $y$ is projected by $P(\theta)$, which depends on $\theta$. It justifies the name “variable projection method” used for the procedure.

Minimization of (4.17b) w.r.t. $\theta$ gives the separable nonlinear least squares estimate $\hat{\theta}_{SNLS}(y)$. Finally, substituting this value in (4.16), we find the estimate $\hat{C}_{SNLS}(y)$.

### 4.3.2. Fitting a Sum of Harmonically Related Sinewaves

Consider the estimation of a sum of harmonically related sinewaves with unknown amplitudes $A_l$, phases $\phi_l$ and frequency $f = \omega/(2\pi)$, from $N$ noisy samples, $k = 1, 2, \ldots, N$,

$$y[k] = \sum_{l=0}^{L} A_l \cos(l\omega_0(k - 1)T_s + \phi_l) + v[k] \quad (4.20)$$

where $T_s$ is the sampling period. Using $\cos(x+y) = \cos(x)\cos(y) - \sin(x)\sin(y)$, (4.20) can be written under the form (4.14) as

$$H[k+1](\theta) = \begin{cases} 
\cos(l\omega_0(k - 1)T_s) & l = 0, 1, \ldots, L \\
\sin(l\omega_0(k - 1)T_s) & l = L + 1, L + 2, \ldots, 2L 
\end{cases} \quad (4.21a)$$

$$C[l+1] = \begin{cases} 
A_l \cos \phi_l & l = 0, 1, \ldots, L \\
-A_l \sin \phi_l & l = L + 1, L + 2, \ldots, 2L 
\end{cases} \quad (4.21b)$$

for $k = 1, 2, \ldots, N$ and with $\theta = f = \omega/(2\pi)$. Applying the variable projection method (4.17) to (4.21), reduces the original nonlinear optimization problem with $2L+2$ model parameters to a problem with a single parameter $f$ (Pintelon and Schoukens, 1996).

### 4.4. Minimizing the Cost Function

In decreasing order of complexity and decreasing convergence speed, the following nonlinear optimization algorithms are discussed (Fletcher, 1991): Newton-Raphson [Section 4.4.1], Gauss-Newton [Section 4.4.2], Levenberg-Marquardt [Section 4.4.3], and gradient descent [Section 4.4.4]. These methods can be fine-tuned by adding a line search [see Section 4.4.5]. Two stopping criteria for
these iterative algorithms are presented in Section 4.4.6. All these methods try to find in an iterative way the global minimum of the cost function (4.3), which can be written as

$$V_{\text{WNLS}}(\theta, y) = \frac{1}{2}e^T(\theta, y)e(\theta, y)$$  \hspace{1cm} (4.22a)

$$e(\theta, y) = W^{1/2} (y - H(\theta))$$  \hspace{1cm} (4.22b)

and where $W^{1/2}$ is a square root of the positive definite matrix $W$. This minimization problem is equivalent to finding the solution of the following set of nonlinear equations

$$V_{\text{WNLS}}^T(\theta, y) = 0 \Rightarrow e^T(\theta, y)e(\theta, y) = 0$$ \hspace{1cm} (4.23)

where $'$ denotes the derivative w.r.t. $\theta$.

### 4.4.1. Newton-Raphson

In numerical analysis, the Newton-Raphson method is an iterative root-finding algorithm for nonlinear functions $F(\theta) = 0$. The $p+1$-th iteration step of this procedure is of the form (Ralston and Rabinowitz, 1984)

$$F'(\theta^{[p]})\Delta \theta^{[p+1]} = -F(\theta^{[p]}) \quad \text{with} \quad \Delta \theta^{[p+1]} = \theta^{[p+1]} - \theta^{[p]}$$ \hspace{1cm} (4.24)

If started “close enough” to a solution, then the convergence of the iterative algorithm (4.24) is quadratic. The important issue of finding an appropriate starting value $\theta^{[0]}$ is addressed in Section 4.5.

Applying (4.24) to the nonlinear function (4.23), gives the Newton-Raphson method for minimizing the cost function (4.22)

$$\begin{align*}
(J^T(\theta^{[p]}), J(\theta^{[p]}), D(\theta^{[p]}, y))\Delta \theta^{[p+1]} & = -J^T(\theta^{[p]})e(\theta^{[p]}, y) \hspace{1cm} (4.25a) \\
J(\theta) & = -W^{1/2}H'(\theta) \hspace{1cm} (4.25b) \\
D_{m,n}(\theta, y) & = -\sum_{k,l=1}^{N} W_{[l,k]} \frac{\partial^2 H_{[k]}(\theta)}{\partial \theta_{[m]} \partial \theta_{[n]}} e_{[l]}(\theta, y) \hspace{1cm} (4.25c)
\end{align*}$$

where $J(\theta)$ is the Jacobian matrix of the residual $e(\theta, y)$ (4.22b). The disadvantages of this method are the calculation of (i) the second order derivatives (4.25c), and (ii) the matrix product $J^T(\theta)J(\theta)$. Hence, the quadratic convergence is at the cost of calculation time and loss in numerical accuracy. However, compared with the other minimization methods, less iteration steps are needed so that the overall computational cost might still be lower. Also the numerical issues are partially accommodated by the iterative nature of the algorithm.
4.4.2. Gauss-Newton

The Gauss-Newton method discards the matrix $D(\theta^{[p]}, y)$ in the Newton-Rapson method (4.25), giving

$$ J^T(\theta^{[p]}) J(\theta^{[p]}) \Delta \theta^{[p+1]} = -J^T(\theta^{[p]}) e(\theta^{[p]}, y) \quad (4.26) $$

This simplification is at the cost of a reduction in convergence speed, which is super linear for Gauss-Newton (Fletcher, 1991). However, since (4.26) is of the same form as (3.3a), the parameter update $\Delta \theta^{[p+1]}$ can be calculated in a numerical reliable way using the SVD or QR decomposition of the (scaled) Jacobian matrix $J(\theta^{[p]})$ [see Section 3.2.2 on page 38]. For example, using $J(\theta) = U(\theta) \Sigma(\theta) V^T(\theta)$, we get

$$ \Delta \theta^{[p+1]} = -V(\theta^{[p]}) \Sigma^{-1}(\theta^{[p]}) U^T(\theta^{[p]}) e(\theta^{[p]}, y) \quad (4.27) $$

4.4.3. Levenberg-Marquardt

Neither the Newton-Raphson nor the Gauss-Newton updates guarantee that the cost function (4.22) will decrease. A technique that ensures decrease of the cost function in each iteration step is the Levenberg-Marquardt method. Via a tunable parameter $\lambda \geq 0$, it decreases the step size of the Gauss-Newton step (4.26) till the cost decreases:

$$ (J^T(\theta^{[p]}) J(\theta^{[p]})) + \lambda^2 I_n \mathbf{a} \Delta \theta^{[p+1]} = -J^T(\theta^{[p]}) e(\theta^{[p]}, y) \quad (4.28) $$

Using the SVD $J(\theta) = U(\theta) \Sigma(\theta) V^T(\theta)$, the Levenberg-Marquardt update (4.28) is calculated in a numerical reliable way as

$$ \Delta \theta^{[p+1]} = -V(\theta^{[p]}) \Sigma^{-1}(\theta^{[p]}) U^T(\theta^{[p]}) e(\theta^{[p]}, y) \quad (4.29a) $$

$$ \Lambda(\theta) = \text{diag}(\frac{\sigma_1}{\sigma_1^2 + \lambda^2}, \ldots, \frac{\sigma_n}{\sigma_n^2 + \lambda^2}) \quad \text{with} \quad \sigma_k = \sigma_k(J(\theta)) \quad (4.29b) $$

where (4.29a) uses $I_n = V^T(\theta^{[p]}) V(\theta^{[p]})$. The following strategy is used for initializing and updating $\lambda$:

1. Start with $\lambda = 0$ and perform Gauss-Newton steps (4.27) so long as the cost function decreases.

2. If the cost function increases, then we keep the previous solution $\theta^{[p]}$, and initialize $\lambda$ as

$$ \lambda = \frac{1}{100} \max_{1 \leq k \leq n_\theta} (\sigma_k(J(\theta^{[p]}))) \quad (4.30) $$

Next, the following steps are executed:
(a) Calculate the Levenberg-Marquardt update (4.29).

(b) If the cost function decreases, then \( \lambda \) is decreased as

\[
\lambda \rightarrow 0.4\lambda \quad (4.31a)
\]

and we go to step (a).

(c) If the cost function increases, then we keep the previous solution \( \theta^{[p]} \), increase \( \lambda \) as

\[
\lambda \rightarrow 10\lambda \quad (4.31b)
\]

and go to step (a).

If \( \lambda^2 I_{n_d} \gg J^T(\theta) J(\theta) \), then the Levenberg-Marquardt method (4.28) simplifies to the gradient method discussed in the next section.

4.4.4. Gradient Descent

The gradient descent method updates the model parameters in the negative direction of the gradient of the cost function

\[
\Delta \theta^{[p+1]} = -\frac{1}{\lambda^2} J^T(\theta^{[p]}) e(\theta^{[p]}, y) \quad (4.32)
\]

where \( \lambda \) can be initialized and updated as in (4.30) and (4.31).

4.4.5. Line Search

The basic idea consists in searching a minimum of the cost function along the direction of the parameter update \( \Delta \theta^{[p+1]} \)

\[
\theta^{[p+1]} = \theta^{[p]} + \gamma \Delta \theta^{[p+1]} \quad (4.33)
\]

Among the many existing algorithms for finding an optimal value for \( \gamma \) (Fletcher, 1991), the following procedure is a good compromise between computational cost and performance. It consists in calculating one additional cost function value with \( \gamma = 0.5 \) in (4.33); those for \( \gamma = 0 \) and \( \gamma = 1 \) being already known:

\[
C(0) = V_{\text{WNLS}}(\theta^{[p]}, y) \quad (4.34a)
\]

\[
C(0.5) = V_{\text{WNLS}}(\theta^{[p]} + 0.5\Delta \theta^{[p+1]}, y) \quad (4.34b)
\]

\[
C(1) = V_{\text{WNLS}}(\theta^{[p+1]}, y) \quad (4.34c)
\]
Figure 4.1: Graphical representation of the line search algorithm. Along the direction of the parameter update, a parameter value halfway between $\theta^{[p]}$ and $\theta^{[p+1]}$ is generated [(4.33) with $\gamma = 0.5$: middle blue ‘×’]. A parabola [green lines] is constructed through the three parameter values [blue ‘×’] and its minimum is calculated [red ‘o’]. The corresponding argument [(4.33) with $\gamma$ given by (4.35)] is the new parameter value. It lies closer to the minimum $\theta = 5$ of the cost function [black lines] than the original values $\theta^{[p]}$ and $\theta^{[p+1]}$.

Next, a quadratic function of $\gamma$ is constructed through the cost function values (4.34), and this function is minimal for [see Figure 4.1 for a graphical representation]

$$
\gamma_{opt} = -\frac{1}{2} \frac{4C(0.5) - C(1) - 3C(0)}{-4C(0.5) + 2C(1) + 2C(0)}
$$

(4.35)

[Exercise: prove (4.35)]. Finally, the parameter update (4.33) is calculated for $\gamma = \gamma_{opt}$, and it is verified whether the corresponding cost function value is smaller than (4.34c). If not, the solution for $\gamma = 1$ is retained [by construction $C(1)$ is smaller than $C(0)$]. Note that the denominator of (4.35) is zero only if the three cost function values lie on a straight line.

### 4.4.6. Stopping Criteria

The iterative procedures are stopped if the relative variation of the model parameters and/or the relative variation of the cost function are smaller than a user defined level:

$$
\frac{\|\theta^{[p+1]} - \theta^{[p]}\|_2}{\|\theta^{[p]}\|_2} \leq \delta
$$

(4.36a)

$$
\frac{|V_{\text{WNLS}}(\theta^{[p+1]}, y) - V_{\text{WNLS}}(\theta^{[p]}, y)|}{V_{\text{WNLS}}(\theta^{[p]}, y)} \leq \epsilon
$$

(4.36b)

or if a maximum number of iterations is reached.
4.5. Generation of Starting Values

There are basically three approaches for generating starting values: (i) use of prior knowledge, (ii) brute force calculations [see Section 4.5.1], and (iii) convex approximation [see Section 4.5.2]. In some physical modeling problems one knows the order of magnitude of one or more physical parameters that can be used as initial guess. Starting values for the unknown parameters must then be found using brute force techniques.

4.5.1. Brute Force Calculations

In this approach a reasonable range for each parameter is discretized in \( N \) samples and the cost function is evaluated for each of the \( n_\theta \) parameters values. As starting value we take the parameter value corresponding to the smallest cost function. Given the exponential growth in cost function evaluations, this method is feasible for a small number of parameters only \( [n_\theta \leq 3]\).

Refinements of this brute force approach exist, where initial values are generated according to some stochastic mechanism. According to the mechanism used, one distinguishes simulated annealing (Nelles, 2001; Press et al., 2007), evolutionary and genetic algorithms (Nelles, 2001). Although very useful, convergence to the global minimum is never guaranteed.

As an example, we take the sum of sinewaves fitting problem (4.20) with \( L = 1, N = 100, f_0 = 0.75f_s/N, A_0 = -0.5, A_1 = 1, \) and \( \phi_1 = \pi/4, \) and where \( v(t) \) is normally distributed with zero mean and standard deviation \( \sigma_v = 0.5. \)

Figure 4.2 shows the true signal and the noisy samples used for estimation [left plot], and the cost function (4.19) as a function of the frequency \( f \) normalized on the DFT resolution \( f_s/N, \) with \( f_s \) the sampling frequency. Beside the global minimum, two local minima can be observed in the interval \([0, 3.75]\), showing the importance of “good enough” starting values. In this case, a scan of the cost function over the DFT frequencies \( k f_s/N, k = 1, 2, \ldots, \) [red ‘x’ in the right plot of Figure 4.17b] is sufficient to avoid the local minima.

4.5.2. Convex Approximation

Sometimes it is possible to make a convex approximation of the modeling problem, so that starting values can be generated using convex optimization techniques (Boyd and Vandenberghe, 2004). These techniques find the global minimum of a convex minimization problem in at most polynomial time. In the remainder of this section we give two examples of nonlinear minimization problems that can be approximated in linear least squares sense.
4.5.2.1. Estimation of the Parameters of an Exponential Function

A first example is the estimation of the parameters of an exponentially decreasing function from noisy measurements.

\[ y(t) = ae^{-t/\tau} + v(t) \]  \hspace{1cm} (4.37)

with \( \tau > 0 \). Using the variable projection method of Section 4.3.1, we get a nonlinear minimization problem (4.19) in \( \tau \). Hence, only a starting value for \( \tau \) is needed. It is obtained as follows. First, taking the logarithm of the absolute value of (4.37) transforms the nonlinear minimization problem into a linear least squares problem in \( \log |a| \) and \( 1/\tau \)

\[ \log |y(t)| \approx \log |a| - \frac{t}{\tau} \]  \hspace{1cm} (4.38)

Note that (4.38) is exact in the noiseless case only. Next, the linear least squares estimate (3.21) of (4.38) is calculated, from which we derive an initial value for \( \tau \). [Question: Is it possible to consistently estimate \( a \) and \( \tau \)? If not why not?]

The whole procedure is illustrated in Figure 4.3 for the case \( a = 1, \tau = 10T_s \), with \( T_s \) the sampling period, and where \( v(t) \) is zero mean Gaussian noise with standard deviation \( \sigma_v = 0.5 \). The parameter values obtained from the linear least squares estimate are \( |\hat{a}| = 0.76 \) and \( \hat{\tau} = 4.5T_s \). It can be seen that the linear least squares estimate of \( \tau \) [red ‘×’ in the right plot] is close enough to the global minimum of the cost function (4.19) to ensure convergence of the nonlinear optimization algorithms.
Figure 4.3: Estimation of the exponential function parameters from noisy observations. Left: true signal [blue] and noisy samples [red]. Right: cost function [blue] (4.17b) as a function of the normalized delay. The red ‘×’ is the value obtained from the linear least squares estimate of (4.38).

4.5.2.2. Curve Fitting using Rational Functions

A second example is the rational approximation of order $R$

$$y(u, \theta) = \frac{\sum_{r=1}^{R+1} \theta[r]u^{r-1}}{\sum_{r=1}^{R+1} \theta[r+R+1]u^{r-1}}$$

(4.39)

of the arctan function over the interval $[0, 5]$

$$V_{\text{NLS}}(\theta, y) = \sum_{k=1}^{N} \left( \text{arctan}(u(k)) - y(u(k), \theta) \right)^2$$

(4.40a)

$$u(k) = \frac{5}{N-1} (k - 1) \text{ with } k = 1, 2, \ldots, N$$

(4.40b)

Multiplying both sides of the equation $\text{arctan}(u(k)) \approx y(u(k), \theta)$ by the denominator of the rational function (4.39), we get

$$\text{arctan}(u(k)) \frac{\sum_{r=1}^{R+1} \theta[r+R+1]u^{r-1}(k)}{\sum_{r=1}^{R+1} \theta[r]u^{r-1}(k)} \approx \sum_{r=1}^{R+1} \theta[r]u^{r-1}(k)$$

(4.41)

This equation is linear-in-the-parameters and the corresponding linear least squares cost function is

$$V_{\text{LLS}}(\theta, y) = \sum_{k=1}^{N} \left( \text{arctan}(u(k)) \sum_{r=1}^{R+1} \theta[r+R+1]u^{r-1}(k) - \sum_{r=1}^{R+1} \theta[r]u^{r-1}(k) \right)^2$$

(4.42)

where $u(k)$ is defined in (4.40b). The solution of (4.42) is used as initial estimate for the nonlinear minimization of (4.40).
Note that the rational approximation problems (4.40) and (4.42) are not well-posed, because the rational function (4.39) remains unchanged if we multiply the numerator and the denominator coefficients with the same non-zero real number \( \lambda \). Hence, a parameter constraint must be added to make the problem identifiable. Typically, one coefficient of the rational function is fixed to one, for example, the denominator coefficient of the highest power of \( u \). A better solution consists in imposing that the 2-norm of the parameter vector is equal to one

\[
\|\theta\|_2 = 1
\]

The linear least squares problem (4.42) subject to constraint (4.43) results in the so-called total least squares (TLS) estimator (Van Huffel and Vandewalle, 1991; Markovsky and Van Huffel, 2007). It is calculated as the right singular vector of the regression matrix \( H \) corresponding to the smallest singular value. Assuming that the singular values are ordered from large to small, we find

\[
H = U\Sigma V^T \rightarrow \hat{\theta}_{\text{TLS}} = V[:,n_\theta]
\]

(Van Huffel and Vandewalle, 1991; Markovsky and Van Huffel, 2007; Pintelon and Schoukens, 2012).

Due to the parameter ambiguity in (4.39), the Jacobian matrix of the residuals in (4.40a)

\[
J[k,r](\theta) = -\frac{\partial y(u(k), \theta)}{\partial \theta[r]} \quad \text{with} \quad k = 1, 2, \ldots, N \quad \text{and} \quad r = 1, 2, \ldots, n_\theta
\]

has rank \( n_\theta - 1 \). Therefore, the Gauss-Newton parameter update (4.27) is replaced by

\[
\Delta \theta[p+1] = -V(\theta[p])\Sigma^+(\theta[p])U^T(\theta[p])e(\theta[p], y)
\]

\[
\Sigma^+ = \text{diag}(\sigma_1^{-1}, \sigma_2^{-1}, \ldots, \sigma_{n_\theta-1}^{-1}, 0)
\]

(4.46a)

(4.46b)

and constraint (4.43) is imposed on the updated parameter

\[
\theta[p+1] = \theta[p] + \Delta \theta[p+1] \rightarrow \frac{\theta[p+1]}{\|\theta[p+1]\|_2}
\]

(4.47)

Following the same lines for the Levenberg-Marquardt method, \( \sigma_{n_\theta} \) is set to zero in (4.29b).

Figure 4.4 shows the result of the rational approximation problems (4.40) and (4.42) for the case \( N = 200 \) and \( R = 7 \). Comparing Figure 4.4 and Figure
$y = \arctan(u)$

Figure 4.4: Rational approximation of the arctan function [left] using a rational function (4.39) of order $R = 7$. Relative approximation error [right] of the total least squares [blue] and nonlinear least squares [red] solutions (4.42) and (4.40).

3.1 on page 42, it can be seen that the relative errors of the rational approximation of order 7 [red line] and the polynomial approximation of order 30 [red line], are the same. However, the rational model has only $2 \times (7 + 1) - 1 = 15$ free model parameters, while the polynomial model $30 + 1 = 31$. It illustrates the importance of the choice of the basis functions used.
4.6. Appendix: Proof of (4.9) and (4.10)

The derivative of the cost function (4.3) w.r.t. \( \theta \) equals

\[
V_{\text{WNLS}}^{\theta T}(\theta, y) = -H^{\theta T}(\theta)W(y - H(\theta))
\]  

(4.48)

Evaluation (4.48) at \( \theta = \theta_0 \), taking into account (4.1), gives

\[
V_{\text{WNLS}}^{\theta T}(\theta_0, y) = -H_0^{\theta T}Wv
\]  

(4.49)

Using (4.7) we find the derivative of the expected value of the cost function

\[
V_{\text{WNLS}}^{\theta T}(\theta) = -H^{\theta T}(\theta)W(H(\theta_0) - H(\theta))
\]  

(4.50)

The derivative of (4.50), evaluated at \( \theta = \theta_0 \), equals

\[
V_{\text{WNLS}}^{\theta T}(\theta_0) = H_0^{\theta T}WH_0^{\theta T}
\]  

(4.51)

Combining (4.49) and (4.51) with (2.55), proves (4.9). Finally, (4.10) follows immediately from \( \text{Cov}(v) = C_v \) and (4.9).
Chapter 5
Maximum Likelihood Method

Abstract: Maximum likelihood estimation is based on the principle of looking for those parameter values that maximize the probability of observing the given data. Contrary to (non)linear least squares, it requires the probability density function of the data. In this chapter we explore the stochastic properties of the maximum likelihood method, and we illustrate the approach on a few examples.

Learning Objectives:

• The concept of maximum likelihood estimation.
• General stochastic properties of the maximum likelihood estimator within the classical framework.
• Construction of a maximum likelihood estimator in practical cases – robustness analysis of the stochastic properties.
• Analysis of maximum likelihood estimators under non-standard conditions.

5.1. Definition of the Maximum Likelihood Estimator

To construct the maximum likelihood estimator, one needs the probability density function \( f_z(z) \) of the measurements \( z \in \mathbb{R}^{N \times 1} \). Given \( f_z(z) \) and the model, one can construct the probability density function (pdf) of the data \( z \), given the true model parameters \( \theta_0 \in \mathbb{R}^{n_0 \times 1} \)

\[
f_{z|\theta_0}(z|\theta_0)
\]

(5.1)

For example, consider noisy measurements \( z = z_0 + v \), with \( z_0 \) the true unknown value and \( v \) the disturbing noise. Assuming that \( z_0 \) can be modeled as

\[
z_0 = h(\theta_0)
\]

(5.2)
The maximum likelihood estimate is by definition the maximizing argument of the likelihood function \( f_{z|\theta}(z|\theta) \).

The conditional probability density function (5.1) equals

\[
f_{z|\theta_0}(z|\theta_0) = f_z(z - h(\theta_0)) \tag{5.3}
\]

The maximum likelihood estimate \( \hat{\theta}_{ML}(z) \) is then defined as the maximizing argument (mode) of the conditional pdf (5.1)

\[
\hat{\theta}_{ML}(z) = \arg \max_{\theta} f_{z|\theta}(z|\theta) \tag{5.4}
\]

where \( f_{z|\theta}(z|\theta) \) is called the likelihood function. Often, the maximization problem (5.4) is numerically ill-defined. Therefore, it is replaced by

\[
\hat{\theta}_{ML}(z) = \arg \min_{\theta} -\log f_{z|\theta}(z|\theta) \tag{5.5}
\]

For Gaussian distributions, (5.5) results in a (non)linear least squares problem [see Section 5.3]. If the derivative of likelihood function exists, then \( \hat{\theta}_{ML}(z) \) is the solution of the following (non)linear set of equations

\[
\left( \frac{\partial - \log f_{z|\theta}(z|\theta)}{\partial \theta} \right)^T = 0 \tag{5.6}
\]

Figure 5.1 shows an example of a non-differentiable likelihood function. By definition, the maximum likelihood estimate is that value of \( \theta \) that maximizes the likelihood.

### 5.2. Asymptotic Properties

The asymptotic \( (N \to \infty) \) properties of the maximum likelihood estimator are analyzed under the following technical conditions (Sorenson, 1980; Caines, 1988)
1. The measurements $z[k]$, $k = 1, 2, \ldots, N$, are independently distributed.

2. In a closed and bounded neighborhood of $\theta_0$, the likelihood function $f_{z|\theta}(z|\theta)$ is a continuous function of $\theta$ with continuous second order derivative.

3. Adding measurements should add information about the model parameters $\theta$

$$
\lim_{N \to \infty} \frac{1}{N} \partial^2 - \log f_{z|\theta}(z|\theta) \quad \text{is of full rank} \quad (5.7)
$$

in a closed and bounded neighborhood of $\theta_0$.

4. The boundaries of the likelihood are independent of $\theta$.

5. The log-likelihood of one measurement has finite variance

$$
\text{var} \left( - \log f_{z[k]}|\theta (z[k]|\theta) \right) \leq c_1 < \infty \quad (5.8)
$$

for $k = 1, 2, \ldots, N$ and with $c_1$ independent of $N$.

6. There exists an $\epsilon > 0$ such that

$$
\mathbb{E} \left\{ |h_r(\theta_0, z[k])|^{2+\epsilon} \right\} \leq c_2 < \infty \quad (5.9a)
$$

$$
h_r(\theta, z[k]) = \frac{\partial - \log f_{z[k]}|\theta (z[k]|\theta)}{\partial \theta^{[r]}} \quad (5.9b)
$$

$$
\lim_{N \to \infty} \frac{N}{\left( \text{var}(g_r(\theta_0, z)) \right)^{1+\epsilon/2}} = 0 \quad (5.9c)
$$

$$
g_r(\theta, z) = \sum_{k=1}^{N} h_r(\theta, z[k]) \quad (5.9d)
$$

for $r = 1, 2, \ldots, n_\theta$ and $k = 1, 2, \ldots, N$, and with $c_2$ independent of $N$.

7. The number of model parameters $n_\theta$ is independent of the amount of data $N$.

Although technical condition 7 seems to be satisfied in an obvious way, an example where it is not fulfilled is given in Section 5.3.3.

### 5.2.1. Consistency

Referring to Section 2.4.1 on page 29, the maximum likelihood estimator is strongly consistent if (i) the expected value of the negative log-likelihood function

$$
V_{\text{ML}}(\theta, z) = - \log f_{z|\theta}(z|\theta) \quad (5.10)
$$
is minimal in the true model parameters \( \theta_0 \), and (ii) the cost function (5.10) converges uniformly with probability one to its expected value

\[
V_{\text{ML}}(\theta) = \mathbb{E}\{ - \log f_{\theta}(z|\theta) \}
\]  

(5.11)

in a closed and bounded neighborhood of \( \theta_0 \). Both conditions are checked in the remainder of this section.

The expected value (5.11) can be elaborated as

\[
V_{\text{ML}}(\theta) = -\int_Z \log f_{\theta}(z|\theta)f_{\theta_0}(z|\theta_0)dz
\]  

(5.12)

where \( f_{\theta_0}(z|\theta_0) = f_z(z) \) is the probability density function of the data \( z \), and with \( Z \) the domain of \( z \). Calculating the derivative of (5.12) w.r.t. \( \theta \) [technical condition 4], gives

\[
V_{\text{ML}}'(\theta) = -\int_Z f'_{\theta}(z|\theta)f_{\theta_0}(z|\theta_0)dz
\]  

(5.13)

Evaluating (5.13) at \( \theta = \theta_0 \), we find

\[
V_{\text{ML}}'(\theta_0) = -\int_Z f'_{\theta_0}(z|\theta_0)dz = -\frac{\partial}{\partial \theta_0} \int_Z f_{\theta_0}(z|\theta_0)dz = -\frac{\partial 1}{\partial \theta_0} = 0
\]  

(5.14)

which proves condition (i).

Using the independence of the measurements [technical condition 1], the negative log-likelihood function (5.10) can be written as

\[
V_{\text{ML}}(\theta, z) = -\log f_{\theta}(z|\theta)
\]

\[
= -\log \prod_{k=1}^{N} f_{z[k]}(z[k]|\theta)
\]

\[
= -\sum_{k=1}^{N} \log f_{z[k]}(z[k]|\theta)
\]  

(5.15)

Under technical condition 5, the strong law of large numbers [Section 2.2.3 on page 15] is applicable to (5.15). It shows that \( V_{\text{ML}}(\theta, z) \) (5.10) converges w.p. 1 to \( V_{\text{ML}}(\theta) \) (5.11). Technical condition 2 ensures that this convergence is uniform in a closed and bounded neighborhood of \( \theta_0 \).
5.2.2. Asymptotic Covariance

Applying the results of Section 2.4.2 on page 29, to the negative likelihood function (5.10), gives

\[ \hat{\theta}_{\text{ML}}(z) - \theta_0 \xrightarrow{\text{w.p.1}} \frac{1}{N} \delta_{\theta,\text{ML}}(z) = -V''_{\text{ML}}(\theta_0) V'_{\text{ML}}(\theta_0, z) \quad (5.16) \]

and

\[ \text{“Cov}(\hat{\theta}_{\text{ML}}(z))” = \text{Cov}(\delta_{\theta,\text{ML}}(z)) \]

\[ = V''_{\text{ML}}^{-1}(\theta_0) E\{V'_{\text{ML}}(\theta_0, z) V'_{\text{ML}}(\theta_0, z)\} V''_{\text{ML}}^{-1}(\theta_0) \quad (5.17a) \]

with

\[ V''_{\text{ML}}(\theta_0) = -E \left\{ \frac{\partial^2 f_{z|\theta_0}(z|\theta_0)}{\partial \theta_0^2} \right\} \quad (5.17b) \]

\[ E\{V'_{\text{ML}}(\theta_0, z) V'_{\text{ML}}(\theta_0, z)\} = E \left\{ \left( \frac{\partial \log f_{z|\theta_0}(z|\theta_0)}{\partial \theta_0} \right) \right\} \]

\[ \frac{\partial V_{\text{ML}}(\theta_0, z)}{\partial \theta_0[r]} = \sum_{k=1}^{N} h_r(\theta_0, z[k]) \quad (5.17d) \]

and where \( h_r(\theta, z[k]) \) is defined in (5.9b). Combining (5.17) with (2.69) on page 34, we finally get

\[ \text{“Cov}(\hat{\theta}_{\text{ML}}(z))” = V''_{\text{ML}}^{-1}(\theta_0) \quad (5.18) \]

where \( V''_{\text{ML}}(\theta_0) \) is defined in (5.17b).

5.2.3. Asymptotic Efficiency

Comparing the Cramér-Rao lower bound (2.15) to (5.18), it follows that the maximum likelihood estimator (5.4) is asymptotically efficient.

5.2.4. Asymptotic Normality

Under technical condition 5.9, the second version of the central limit theorem [see Section 2.2.5 on page 16] can be applied to the sum (5.17d), which shows that \( \delta_{\theta,\text{ML}}(z) \) (5.16) is asymptotically normally distributed [\( V''_{\text{ML}}(\theta_0) \) does not depend on \( z \)].
5.2.5. Robustness

No general statements about the robustness of the maximum likelihood estimator (5.4) can be made. However, for some specific cases a robustness analysis is possible [see Section 5.3].

5.2.6. Invariance Principle

Consider any function \( \psi = g(\theta) \in \mathbb{R}^{n_\psi} \) with \( n_\psi \leq n_\theta \). If \( \hat{\theta}_{\text{ML}} \) is the maximum likelihood estimate of \( \theta \), then the maximum likelihood estimate \( \hat{\psi}_{\text{ML}} \) of \( \psi \) is given by

\[
\hat{\psi}_{\text{ML}} = g(\hat{\theta}_{\text{ML}})
\]  

(5.19)

Result (5.19) follows from the equivalence

\[
\max_{\theta} f_{z|\theta}(z|\theta) = \max_{\psi \text{ s.t. } g(\theta) = \psi} \max_{\theta} f_{z|\theta}(z|\theta)
\]  

(5.20)

5.3. Examples

Three examples of Gaussian maximum likelihood estimators are discussed: (i) the sample mean and sample variance of \( N \) independent samples [Section 5.3.1], (ii) the estimation of the resistor value from noisy DC current and voltage measurements [Section 5.3.2], and (iii) the estimation of the slope of a straight line through the origin from noisy abscissa and ordinate measurements [Section 5.3.3]. The first two fit within the classical maximum likelihood framework, while the third does not. For each example, the robustness w.r.t. the basic assumptions made to construct the estimator is shown.

5.3.1. Sample Mean and Sample Variance

Suppose that \( z_{[k]}, k = 1, 2, \ldots, N \) are independent samples from a normal distribution with unknown mean \( \mu_0 \) and variance \( \sigma_0^2 \). The likelihood function is

\[
f_{z|\mu,\sigma^2}(z|\mu,\sigma^2) = \prod_{k=1}^{N} f_{z_{[k]}|\mu,\sigma^2}(z_{[k]}|\mu,\sigma^2)
\]

\[
= \frac{1}{(2\pi\sigma^2)^{N/2}} e^{-\sum_{k=1}^{N} \frac{(z_{[k]}-\mu)^2}{2\sigma^2}}
\]  

(5.21)

with corresponding negative log likelihood (5.10)

\[
V_{\text{ML}}(\mu,\sigma^2, z) = \frac{N}{2} \log(2\pi\sigma^2) + \frac{1}{2\sigma^2} \sum_{k=1}^{N} (z_{[k]} - \mu)^2
\]  

(5.22)
The maximum likelihood (ML) estimates \( \hat{\mu}_{\text{ML}}(z) \) and \( \hat{\sigma}^2_{\text{ML}}(z) \) are the solutions of

\[
\frac{\partial V_{\text{ML}}(\mu, \sigma^2, z)}{\partial \mu} = 0 \quad \Rightarrow \quad -\frac{1}{\sigma^2} \sum_{k=1}^{N} (z[k] - \mu) = 0 \quad (5.23a)
\]

\[
\frac{\partial V_{\text{ML}}(\mu, \sigma^2, z)}{\partial \sigma^2} = 0 \quad \Rightarrow \quad \frac{N}{2\sigma^2} - \frac{1}{2\sigma^4} \sum_{k=1}^{N} (z[k] - \mu)^2 = 0 \quad (5.23b)
\]

and we find

\[
\hat{\mu}_{\text{ML}}(z) = \frac{1}{N} \sum_{k=1}^{N} z[k] \quad (5.24a)
\]

\[
\hat{\sigma}^2_{\text{ML}}(z) = \frac{1}{N} \sum_{k=1}^{N} (z[k] - \hat{\mu}_{\text{ML}}(z))^2 \quad (5.24b)
\]

In the remainder of this section we analyze the stochastic properties of the ML estimates (5.24).

### 5.3.1.1. Asymptotic Properties ML Estimates (5.24)

Technical condition 1 is satisfied by assumption. From (5.21) and (5.22) it follows that technical conditions 1, 2, 4 and 5 are fulfilled. Two parameters are estimated so that technical condition 7 is also met.

The second order derivative w.r.t. \( \theta = [\mu \sigma^2]^T \) of the negative log likelihood (5.22)

\[
\frac{1}{N} \frac{\partial^2 V_{\text{ML}}(\mu, \sigma^2, z)}{\partial \theta^2} = \left[ \frac{1}{N\sigma^2} \sum_{k=1}^{N} (z[k] - \mu) \quad \frac{1}{N\sigma^4} \sum_{k=1}^{N} (z[k] - \mu)^2 - \frac{\sigma^2}{2} \right] \quad (5.25)
\]

has asymptotically \( (N \to \infty) \) full rank in the neighborhood of the true values \( \mu_0 \) and \( \sigma_0^2 \). Hence, technical condition 3 is satisfied.

Finally, the left hand sides of (5.23) define the functions \( h_r(\theta, z[k]) \) in (5.9b)

\[
h_1(\mu, \sigma^2, z[k]) = -\frac{1}{\sigma^2} (z[k] - \mu) \quad (5.26a)
\]

\[
h_2(\mu, \sigma^2, z[k]) = \frac{1}{2\sigma^2} - \frac{1}{2\sigma^4} (z[k] - \mu)^2 \quad (5.26b)
\]

These functions have finite moments of any order, so that (5.9a) is fulfilled. Since the variance of \( g_r(\theta_0, z) \) is \( \sum_{k=1}^{N} h_r(\theta_0, z[k]) \) is an \( O(N) \), (5.9c) is valid and, hence, technical condition 6 is also satisfied.
We conclude that the ML estimates (5.24) are strongly consistent, asymptotically normally distributed, and asymptotically efficient. The asymptotic covariance of the estimates is found by calculating the inverse of the expected value of $N$ times (5.25)

\[
{\text{Cov}}(\hat{\theta}_{\text{ML}}(z)) = V_{\text{ML}}^{-1}(\mu_0, \sigma_0^2) = \begin{bmatrix}
\frac{\sigma_0^2}{N} & 0 \\
0 & \frac{2\sigma_0^4}{N^2}
\end{bmatrix}
\] (5.27)

Note that the variance of both the sample mean and sample variance decrease to zero as an $O(N^{-1})$, and are asymptotically uncorrelated.

5.3.1.2. Robustness ML Estimates (5.24)

The cost function (5.22) satisfies the conditions of the strong law of large numbers (2.8a), which does not make use of the probability density function of the data. Hence, the strong consistency of the ML estimates (5.24) remains valid for non-Gaussian i.i.d. measurements with finite second order moments. If the measurements are correlated, then a mixing condition of order 4 is needed \[Question: Why of order 4? Hint: use the results of Section 2.2.3\]. The ML estimates are then still weakly consistent.

For non-Gaussian i.i.d. measurements with finite moments of order 4, the functions (5.26) satisfy the conditions of the first version of the central limit theorem [see Section 2.2.5 on page 16]. Hence, the asymptotic normality of the ML estimates (5.24) remains valid. If the measurements are correlated and mixing of order infinity, then, according to the third version of the central limit theorem [Section 2.2.5], the ML estimates (5.24) are still asymptotically normally distributed.

For non-Gaussian distributed measurements, the asymptotic efficiency property is lost, and the asymptotic covariance is no longer given by (5.27), but by the general expression (5.17a). To calculate the latter, one needs – besides the variance – also the knowledge of the third and fourth order moments of the measurements.

5.3.1.3. Bias ML Estimates (5.24)

For finite values of $N$, the expected value of the ML estimates (5.24) equals

\[
E\{\hat{\mu}_{\text{ML}}(z)\} = \mu_0
\] (5.28)

\[
E\{\hat{\sigma}_{\text{ML}}^2(z)\} = \frac{N - 1}{N} \sigma_0^2
\] (5.29)
It follows that the ML estimate of the mean is unbiased while that of the variance is biased for finite values of $N$. This finite sample bias is removed as

$$
\hat{\sigma}^2(z) = \frac{1}{N-1} \sum_{k=1}^{N} (z[k] - \hat{\mu}_{ML}(z))^2
$$

(5.30)

### 5.3.2. Estimation Resistor Value from DC Measurements

Consider again the estimation of the resistor value from noisy DC current and voltage measurements [see Section 1.2.1 on page 3]. The three true model parameters $\theta_0 = [u_0 \ i_0 \ R_0]^T$ are related by Ohm’s law

$$
u_0 = R_0 i_0
$$

(5.31)

Hence, in total there are two free model parameters. Since all current and voltage measurements are independently distributed, the likelihood function $f_{z|R,i}(z|R,i)$, where $z \in \mathbb{R}^{2N \times 1}$ is the vector of the DC current and DC voltage measurements (2.37), equals the product of the pdfs of each measurement separately (2.38). Using the Gaussian prior of the measurements, the corresponding negative log likelihood equals within a $\theta$-independent term,

$$
V_{ML}(R,i,z) = \frac{1}{2\sigma_u^2} \sum_{k=1}^{N} (u(k) - Ri)^2 + \frac{1}{2\sigma_i^2} \sum_{k=1}^{N} (i(k) - i)^2
$$

(5.32)

The maximum likelihood estimates $\hat{R}_{ML}(z)$ and $\hat{i}_{ML}(z)$ are the solutions of

$$
\frac{\partial V_{ML}(R,i,z)}{\partial R} = 0 \Rightarrow \frac{i}{\sigma_u^2} \sum_{k=1}^{N} (u(k) - Ri) = 0
$$

(5.33a)

$$
\frac{\partial V_{ML}(R,i,z)}{\partial i} = 0 \Rightarrow \frac{R}{\sigma_u^2} \sum_{k=1}^{N} (u(k) - Ri) - \frac{1}{\sigma_i^2} \sum_{k=1}^{N} (i(k) - i) = 0
$$

(5.33b)

We find

$$
\hat{i}_{ML}(z) = \frac{1}{N} \sum_{k=1}^{N} i(k)
$$

(5.34a)

$$
\hat{R}_{ML}(z) = \frac{1}{N} \frac{\sum_{k=1}^{N} u(k)}{\sum_{k=1}^{N} i(k)}
$$

(5.34b)

Note that the ML estimate of the DC current is unbiased, while the expected value of the resistor estimate (5.34b) does not exist for normally distributed current measurements.
5.3.2.1. Asymptotic Properties ML Estimates (5.34)

Following the same lines of Section 5.3.1.1, it can easily be verified that technical conditions 1–7 are all fulfilled. Hence, the ML estimates (5.34) are strongly consistent, asymptotically normally distributed and asymptotically efficient. Finally, the asymptotic covariance (5.18) is given by (2.42).

5.3.2.2. Robustness ML Estimates (5.34)

The robustness of the ML estimate of the resistor value (5.34b) has been discussed in Section 2.3.5 on page 27 and further. A same reasoning can be made for the ML estimate of the current (5.34a). It follows that the consistency and asymptotic normality properties of the Gaussian ML estimates (5.34) are robust w.r.t. several of the basic assumptions made to construct the estimator.

5.3.3. Estimation of the Slope of a Straight Line

Consider again the estimation of the slope of a straight line through the origin (1.9) from noisy abscissa $u(k)$ and ordinate $y(k)$ samples, $k = 1, 2, \ldots, N$ [see Section 1.3 on page 8]. The $2N + 1$ true model parameters, the abscissas $u_0(k)$, the ordinates $y_0(k)$, $k = 1, 2, \ldots, N$, and the slope $\theta_0$ are related as

$$y_0(k) = \theta_0 u_0(k)$$  \hspace{1cm} (5.35)

Hence, the number of free model parameters equals the total number of model parameters minus the number of constraints (5.35), giving

$$2N + 1 - N = N + 1$$ \hspace{1cm} (5.36)

Note that the number of free model parameters (5.36) increases with the amount of data $N$, so that technical condition 7 is not satisfied.

Assuming that the measurements are independent and normally distributed as in (1.11), the likelihood function is given by

$$f_z(\theta, u_p|z, \theta, u_p) = \prod_{k=1}^{N} \frac{1}{2\pi \sigma_y(k) \sigma_u(k)} e^{-\frac{(y(k) - \theta u_p(k))^2}{2\sigma_y^2(k)} - \frac{(u(k) - u_p(k))^2}{2\sigma_u^2(k)}}$$  \hspace{1cm} (5.37)

where $z$ is the $2N \times 1$ vector of noisy abscissa and ordinate measurements, and $u_p$ the $N \times 1$ vector of the unknown abscissa values. Within a parameter independent constant, the corresponding negative log likelihood equals

$$V_{ML}(\theta, u_p, z) = \sum_{k=1}^{N} \frac{(y(k) - \theta u_p(k))^2}{2\sigma_y^2(k)} + \sum_{k=1}^{N} \frac{(u(k) - u_p(k))^2}{2\sigma_u^2(k)}$$  \hspace{1cm} (5.38)
Since the cost function (5.38) is a quadratic function of the parameters \( u_p(k) \), \( k = 1, 2, \ldots, N \), they can analytically be eliminated. We find

\[
\frac{\partial V_{ML}(\theta, u_p, z)}{\partial u_p(k)} = 0 \quad \Rightarrow \quad u_p(k) = \frac{\theta \sigma_y^2(k)y(k) + \sigma_y^2(k)u(k)}{\theta^2 \sigma_y^2(k) + \sigma_y^2(k)}
\]  

(5.39)

Substituting (5.39) in (5.38), we get

\[
V_{ML}(\theta, z) = \frac{1}{2} \sum_{k=1}^{N} \frac{(y(k) - \theta u(k))^2}{\sigma_y^2(k) + \theta^2 \sigma_u^2(k)}
\]  

(5.40)

which is a nonlinear least squares problem in \( \theta \) with noisy regression matrix. Starting values for the nonlinear optimization

\[
\hat{\theta}_{ML}(z) = \arg \min_{\theta} V_{ML}(\theta, z)
\]  

(5.41)

are generated via the convex approximation (1.13).

Note that the denominator of each term in the cost function (5.40) is exactly equal to the variance of the equation error \( y(k) - \theta u(k) \). Hence, the maximum likelihood solution automatically emphasizes high signal-to-noise ratio (SNR) measurements and de-emphasizes low SNR data.

Since the nonlinear least squares problem is not a standard maximum likelihood problem (the number of parameters increases with the amount of data), nor a standard nonlinear least squares problem (the regression matrix is noisy), we cannot rely on, respectively, Sections 5.2 and 4.2 for the asymptotic properties. Therefore, the stochastic properties are analyzed following the lines of Section 2.4 on page 28.

5.3.3.1. Strong Consistency ML Estimate (5.41)

Taking into account that the denominator of each term in the cost function (5.40) equals the variance of the equation error, the expected value of the cost is readily found

\[
V_{ML}(\theta) = \mathbb{E}\{V_{ML}(\theta, z)\}
\]

\[
= \frac{N}{2} + \frac{1}{2} \sum_{k=1}^{N} \frac{(y_0(k) - \theta u_0(k))^2}{\sigma_y^2(k) + \theta^2 \sigma_u^2(k)}
\]  

(5.42)

where the term \( N/2 \) in (5.42) originates from the noise on the data. Clearly, the expected value of the cost function (5.42) is minimal in the true value of the slope \( \theta_0 \).
Applying the strong law of large numbers for i.i.d. random variables [see Section 2.2.3 on page 15] to the cost function (5.40) \[ \text{Question: Why can we do that?} \], shows that it converges w.p. 1 to its expected value. This convergence is uniform in \( \theta \) because the cost function (5.40) and its derivative w.r.t. \( \theta \) are continuous functions in a closed and bounded neighborhood of \( \theta_0 \).

We conclude that the ML estimate (5.41) is strongly consistent

\[
\lim_{N \to \infty} \hat{\theta}_{\text{ML}}(z) = \theta_0
\]  

(5.43)

Note, however, that the ML estimates of the abscissa values [see (5.39)]

\[
\hat{u}_{\text{ML}}(k) = \frac{\hat{\theta}_{\text{ML}}(z)\sigma_u^2(k)y(k) + \sigma_y^2(k)u(k)}{\theta_{\text{ML}}^2(z)\sigma_u^2(k) + \sigma_y^2(k)}
\]  

(5.44)

are not consistent. The reason for this is that new measurements give no information about previous abscissa values.

5.3.3.2. Asymptotic Variance ML Estimate (5.41)

Direct application of (2.55) and (2.56) to (5.41), gives

\[
\hat{\theta}_{\text{ML}}(z) - \theta_0 \xrightarrow{w.p.1} \delta_{\theta,\text{ML}}(z) = -V''^{-1}_{\text{ML}}(\theta_0)V'_{\text{ML}}(\theta_0, z)
\]  

(5.45)

and

\[
\text{var}(\hat{\theta}_{\text{ML}}(z)) = \text{var}(\delta_{\theta,\text{ML}}(z)) = V''^{-1}_{\text{ML}}(\theta_0)E\{V'^2_{\text{ML}}(\theta_0, z)\}V''^{-1}_{\text{ML}}(\theta_0)
\]  

(5.45a)

where

\[
V''_{\text{ML}}(\theta_0) = \sum_{k=1}^{N} \frac{u_0^2(k)}{\theta_0^2\sigma_u^2(k) + \sigma_y^2(k)}
\]  

(5.45b)

\[
E\{V'^2_{\text{ML}}(\theta_0, z)\} \geq V''_{\text{ML}}(\theta_0)
\]  

(5.45c)

Equality in (5.45c) is obtained if and only if either \( \sigma_u(k) = 0 \) or \( \sigma_y(k) = 0 \) for \( k = 1, 2, \ldots, N \) [proof: see Pintelon and Hong, 2007].

5.3.3.3. Asymptotic Normality ML Estimate (5.41)

The derivative of the cost function (5.40) w.r.t. \( \theta \), evaluated at \( \theta_0 \), is a sum of independently distributed random variables with existing moments of any order. Hence, the second version of the central limit theorem [see Section 2.2.5 on page 16] can be applied to \( V'_{\text{ML}}(\theta_0, z) \) in (5.45), which proves the asymptotic normality of the ML estimate (5.41).
5.3.3.4. Asymptotic Inefficiency ML Estimate (5.41)

The Fisher information matrix corresponding to the likelihood function (5.37) is given by

\[
Fi(\psi_0) = -\frac{\partial^2 \log \mathbb{E}\{f_z|\theta_0, u_0(z|\theta_0, u_0)\}}{\partial \psi_0^2} \tag{5.47a}
\]

where \( \psi_0 = [\theta, u_p^T]^T \). Using (5.38), we find for (5.47a)

\[
- \frac{\partial^2 \mathbb{E}\{f_z|\theta_0, u_0(z|\theta_0, u_0)\}}{\partial \theta_0^2} = \sum_{k=1}^{N} \frac{u_0^2(k)}{\sigma_y^2(k)} \tag{5.47b}
\]

\[
- \frac{\partial^2 \mathbb{E}\{f_z|\theta_0, u_0(z|\theta_0, u_0)\}}{\partial u_0(k)^2} = \frac{\theta^2 \sigma_y^2(k) + \sigma_u^2(k)}{\sigma_u^2(k) \sigma_y^2(k)} \tag{5.47c}
\]

\[
- \frac{\partial^2 \mathbb{E}\{f_z|\theta_0, u_0(z|\theta_0, u_0)\}}{\partial u_0(k) \partial u_0(l)} = 0 \tag{5.47d}
\]

\[
- \frac{\partial^2 \mathbb{E}\{f_z|\theta_0, u_0(z|\theta_0, u_0)\}}{\partial \theta_0 \partial u_0(k)} = \frac{\theta_0 u_0(k)}{\sigma_y^2(k)} \tag{5.47e}
\]

Using the block matrix inversion lemma, the inverse of the Fisher information matrix (5.47a) can be written as

\[
Fi^{-1}(\psi_0) = \begin{bmatrix}
Fi^{-1}(\theta_0) & \cdots \\
\vdots & \ddots
\end{bmatrix} \tag{5.48a}
\]

where \( Fi(\theta_0) \) is the Fisher information of the slope parameter

\[
Fi(\theta_0) = \sum_{k=1}^{N} \frac{u_0^2(k)}{\theta^2 \sigma_u^2(k) + \sigma_y^2(k)} \tag{5.48b}
\]

[proof: see Appendix 5.4]. Comparing \( Fi^{-1}(\theta_0) \) with (5.46), it can be concluded that the ML estimate (5.41) does not reach the Cramér-Rao lower bound and, hence, is asymptotically inefficient.

5.3.3.5. Robustness ML Estimate (5.41)

The cost function (5.40) satisfies the conditions of the strong law of large numbers (2.8a), which does not make use of the probability density function of the data. Hence, the strong consistency of the ML estimate (5.43) remains valid for non-Gaussian independent measurements with finite second order moments. If the measurements are correlated, then a mixing condition of order 4 is needed. The ML estimates are then still weakly consistent.
For non-Gaussian independent measurements with finite moments of order \(4 + \epsilon\), with \(\epsilon > 0\), the derivative of the cost function \(V_{\text{ML}}'(\theta_0, z)\) in (5.45) satisfies the conditions of the second version of the central limit theorem [see Section 2.2.5 on page 16]. Hence, the asymptotic normality of the ML estimate of the slope remains valid. If the measurements are correlated and mixing of order infinity, then, according to the third version of the central limit theorem [Section 2.2.5], (5.41) is still asymptotically normally distributed.

5.3.3.6. Summary

Although technical condition 7 of the classical ML framework is not met, the ML estimate (5.41) of the slope parameter is still strongly consistent and asymptotically normally distributed. What is lost compared with the classical framework is the asymptotic efficiency. In addition the ML estimates (5.44) of the abscissas are inconsistent.

The consistency and asymptotic normality properties of Gaussian ML estimator (5.41) are robust w.r.t. several of the assumptions made to construct the estimator.
5.4. Appendix: Proof of (5.48)

The block matrix inversion of a symmetric matrix equals

\[
\begin{bmatrix}
A & C^T \\
C & B
\end{bmatrix}^{-1} = \begin{bmatrix}
F^{-1} & -F^{-1}G^T \\
-GF^{-1} & B^{-1} + GF^{-1}G^T
\end{bmatrix} \tag{5.49a}
\]

\[F = A - C^TB^{-1}C\tag{5.49b}\]

\[G = B^{-1}C\tag{5.49c}\]

Applying (5.49) to the Fisher information matrix (5.47), where

\[A = \sum_{k=1}^{N} \frac{u_0^2(k)}{\sigma_y^2(k)}\]

\[B = \text{diag}(\ldots \frac{\theta_0^2 \sigma_y^2(k) + \sigma_y^2(k)}{\sigma_u^2(k) \sigma_y^2(k)} \ldots )\]

\[C[k] = \frac{\theta_0 u_0(k)}{\sigma_y^2(k)}\]

shows that

\[F = \sum_{k=1}^{N} \frac{u_0^2(k)}{\theta_0^2 \sigma_u^2(k) + \sigma_y^2(k)}\]

which proves (5.48).
Chapter 6
Bayesian Approach

Abstract: Sometimes, uncertain prior knowledge about the true model parameters is available. Using the Bayesian framework, this uncertain information is included during estimation. It requires the knowledge of the probability density function of the data and of the uncertain model parameters. As a special case, we handle in more detail the Gaussian prior on the model parameters leading to the concept of Gaussian process modeling. Within this modeling framework, partial prior knowledge about the model parameters can be incorporated by parametrizing the covariance matrix of the Gaussian process. These (hyper)parameters are optimized via a marginal likelihood function.

Learning Objectives:

- The concept of Bayesian estimation and its properties.
- Link between Bayesian estimation and regularized least squares.
- Gaussian process modeling: Bayesian modeling framework with Gaussian prior and parametrized covariance.

6.1. Definition of the Bayesian Estimator

In the Bayesian modeling framework, prior knowledge about (some of) the model parameters $\theta \in \mathbb{R}^{n_{\theta}}$ to be estimated is available under the form of a probability density function (pdf) $f_{\theta}(\theta)$. Using this information together with the pdf $f_z(z)$ of the data $z$, the (posterior) conditional PDF $f_{\theta|z}(\theta|z)$ of the model parameters $\theta$, given the data $z$, is constructed via Bayes’ rule

$$f_{\theta|z}(\theta|z) = \frac{f_z(z|\theta)f_{\theta}(\theta)}{f_z(z)} \quad (6.1a)$$

$$= \frac{f_{z,\theta}(z,\theta)}{f_z(z)} \quad (6.1b)$$

95
where the likelihood function \( f_z(\theta | z) \) in (6.1a) is derived from \( f_z(z) \) and the model [see Section 5.1 on page 79], and with \( f_{z,\theta}(z, \theta) \) the joint distribution of the data \( z \) and the model parameters \( \theta \). Different estimators \( \hat{\theta}(z) \) can be defined from (6.1), for example, the posterior mean, the posterior median and the posterior mode

\[
\text{posterior mean} : \quad \hat{\theta}(z) = \mathbb{E}\{\theta | z\} = \int_{-\infty}^{+\infty} \theta f_{\theta|z}(\theta | z) d\theta \tag{6.2a}
\]

\[
\text{posterior median} : \quad \int_{-\infty}^{\hat{\theta}(z)} f_{\theta|z}(\theta | z) d\theta = \frac{1}{2} \tag{6.2b}
\]

\[
\text{posterior mode} : \quad \hat{\theta}(z) = \arg \max_{\theta} f_{\theta|z}(\theta | z) \tag{6.2c}
\]

Note that for Gaussian measurements \( z \), and a Gaussian prior on \( \theta \), the three estimators in (6.2) coincide.

The posterior mode is called the Bayes estimator. Combining (6.1) with (6.2c) it can be written as

\[
\hat{\theta}_{\text{Bayes}}(z) = \arg \max_{\theta} f_{z|\theta}(z | \theta) f_{\theta}(\theta) = \arg \min_{\theta} - \log \left( f_{z|\theta}(z | \theta) f_{\theta}(\theta) \right) \tag{6.3a}
\]

\[
= \arg \min_{\theta} - \log f_{z,\theta}(z, \theta) \tag{6.3b}
\]

because \( f_z(z) \) is independent of the model parameters \( \theta \).

### 6.2. Asymptotic Properties

From (6.3a) it is seen that the Bayes estimator combines prior knowledge of the model parameters with information from data. When increasing the amount of data, the data information becomes more important than the prior knowledge. Therefore, the Bayes estimator (6.3) converges for \( N \to \infty \) to the maximum likelihood estimator (5.5):

\[
\hat{\theta}_{\text{Bayes}}(z) \xrightarrow{\text{in stoch. sense}} \hat{\theta}_{\text{ML}}(z) \tag{6.4}
\]

Hence, \( \hat{\theta}_{\text{Bayes}}(z) \) (6.3) inherits all asymptotic properties of \( \hat{\theta}_{\text{ML}}(z) \) (5.5) discussed in Chapter 5. However, the finite sample behavior of both estimators can be quite different. This is especially the case if the amount of data \( N \) is not much larger than the number of parameters \( n_\theta \) to be estimated.
6.3. Examples

Two examples of the Bayesian modeling framework are given. In the first example, the estimation of the slope of a straight line through the origin, the impact of the prior knowledge and the data on the Bayes estimate is illustrated. The second example establishes the connection between regularized linear least squares and Bayesian linear regression.

6.3.1. Estimation of the Slope of a Straight Line

We retake the example of estimating the slope of a straight line through the origin from noisy abscissa \(u(k)\) and ordinate \(y(k)\) measurements, \(k = 1, 2, \ldots, N\) [see Section 5.3.3 on page 88]. Furthermore, we assume that the prior distribution of the slope parameter is Gaussian with mean value \(\mu_\theta\) and standard deviation \(\sigma_\theta\)

\[
f_\theta(\theta) = \frac{1}{\sqrt{2\pi\sigma_\theta^2}} e^{-\frac{(\theta-\mu_\theta)^2}{2\sigma_\theta^2}} \quad (6.5)
\]

Since no prior knowledge is available for the abscissa values, the Bayes cost function to be minimized w.r.t. \(\theta\) and \(u_p\), equals

\[
V_{\text{Bayes}}(\theta, u_p, z) = -\log f_{z|\theta, u_p}(z|\theta, u_p) - \log f_\theta(\theta) = \frac{N}{2} \sum_{k=1}^{N} \frac{(y(k) - \theta u_p(k))^2}{2\sigma_y^2(k)} + \frac{N}{2} \sum_{k=1}^{N} \frac{(u(k) - u_p(k))^2}{2\sigma_u^2(k)} + \frac{(\theta - \mu_\theta)^2}{2\sigma_\theta^2} + C_1 \quad (6.6)
\]

where \(C_1\) is a \(\theta\)-independent constant. Proceeding in the same way as in Section 5.3.3, the abscissa values \(u_p(k)\) are eliminated analytically, giving

\[
V_{\text{Bayes}}(\theta, z) = \frac{1}{2} \sum_{k=1}^{N} \frac{(y(k) - \theta u(k))^2}{\sigma_y^2(k) + \theta^2 \sigma_u^2(k)} + \frac{(\theta - \mu_\theta)^2}{2\sigma_\theta^2} \quad (6.7)
\]

Figure 6.1 shows the Bayes cost function (6.7) for different Gaussian priors (6.5), and using the same numerical values as in Section 1.3 on page 8. It can be seen that the impact of the prior knowledge increases as \(\sigma_\theta\) decreases. A wrong mean prior value \(\mu_\theta\) introduces a bias that diminishes as \(\sigma_\theta\) increases. Similar results are obtained if \(N\) is changed instead of \(\sigma_\theta\) [increasing/decreasing \(N\) is equivalent to increasing/decreasing \(\sigma_\theta\)].
Figure 6.1: Bayes cost function (6.7) for different Gaussian priors on the slope parameter $\theta$: $\sigma_\theta = \infty$ [black lines], $\sigma_\theta = 0.1$ [blue lines], and $\sigma_\theta = 0.01$ [red lines]. Top row: correct prior mean $\mu_\theta = 1$. Bottom row: wrong prior mean $\mu_\theta = 1.1$. Right column: zoom around the global minima of the cost functions in the left column.

If the abscissas are known, $\sigma_u(k) = 0$ for $k = 1, 2, \ldots, N$, then (6.7) can be solved analytically. We find

$$\hat{\theta}_{\text{Bayes}}(z) = \frac{\sigma_\theta^2 \sum_{k=1}^{N} y(k) u(k) + \sigma_y^2 \mu_\theta}{\sigma_\theta^2 \sum_{k=1}^{N} u^2(k) + \sigma_y^2} \quad (6.8)$$

The following special cases are worth mentioning

$$\lim_{\sigma_y \to \infty} \hat{\theta}_{\text{Bayes}}(z) = \lim_{\sigma_\theta \to 0} \hat{\theta}_{\text{Bayes}}(z) = \mu_\theta \quad (6.9)$$

$$\lim_{\sigma_\theta \to \infty} \hat{\theta}_{\text{Bayes}}(z) = \frac{\sum_{k=1}^{N} y(k) u(k)}{\sum_{k=1}^{N} u^2(k)} \quad (6.10)$$

It illustrates that if the signal-to-noise ratio of the measurements is very poor ($\sigma_y \to \infty$) or the prior knowledge is very precise ($\sigma_\theta \to 0$), then the Bayes estimate is dominated by the prior knowledge [see (6.9)]. However, if the prior knowledge is very imprecise ($\sigma_\theta \to \infty$) or a lot of measurements are available ($N \gg 1$), then (6.8) is mainly determined by the measurements [see (6.10)].
6.3.2. Bayesian Linear Regression

Consider the linear regression problem

\[
y = H\theta + v
\]

(6.11)

where the regression matrix \( H \in \mathbb{R}^{N \times n_\theta} \) is noise free, and where the noise \( v \in \mathbb{R}^N \) is normally distributed with zero mean and covariance \( \sigma^2 I_N \)

\[
f_v(v) = \frac{1}{\sqrt{(2\pi)^N N\sigma^2}} e^{-\frac{v^T v}{2\sigma^2}}
\]

(6.12)

A Gaussian prior with zero mean and covariance \( C_\theta \) is imposed on the model parameters \( \theta \)

\[
f_\theta(\theta) = \frac{1}{\sqrt{2\pi\det C_\theta}} e^{-\frac{1}{2} \theta^T C_\theta^{-1} \theta}
\]

(6.13)

In this case, the three estimators in (6.2) coincide. Therefore, the Bayes estimate (6.3) will be calculated via the posterior mean (6.2a), using the following key property of Gaussian distributions.

Let \( x_1 \) and \( x_2 \) be jointly Gaussian distributed with mean and covariance

\[
\begin{bmatrix} x_1 \\ x_2 \end{bmatrix} 
\begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix} 
\begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{12}^T & \Sigma_{22} \end{bmatrix}
\]

(6.14a)

If \( \Sigma_{22} \) is regular, then, \( x_1 \) given \( x_2 \), denoted as \( x_1|x_2 \), is normally distributed with mean and covariance

\[
\begin{align*}
E\{x_1|x_2\} &= \mu_1 + \Sigma_{12} \Sigma_{22}^{-1} (x_2 - \mu_2) \\
\text{Cov}(x_1|x_2) &= \Sigma_{11} - \Sigma_{12} \Sigma_{22}^{-1} \Sigma_{12}^T
\end{align*}
\]

(6.15a)

[proof: see Appendix 6.5 on page 106].

The posterior distribution \( f_{\theta|z}(\theta|z) \), can be constructed from the joint pdf of the data \( y \) and the model parameters \( \theta \) [see (6.1b)]. From (6.11), (6.12) and (6.13), it follows that \( y \) and \( \theta \) are jointly Gaussian distributed with zero mean value and covariance

\[
\begin{bmatrix} \theta \\ y \end{bmatrix} 
\begin{bmatrix} C_\theta & C_\theta H^T \\ H C_\theta & H C_\theta H^T + \sigma^2 I_N \end{bmatrix}
\]

(6.16)
Using (6.15), we immediately find

\[ E\{\theta|y\} = C_\theta H^T (HC_\theta H^T + \sigma^2 I_N)^{-1} y \]  
(6.17a)

\[ = (C_\theta H^T H + \sigma^2 I_{n_\theta})^{-1} C_\theta H^T y \]  
(6.17b)

\[ \text{Cov}(\theta|y) = C_\theta - C_\theta H^T (HC_\theta H^T + \sigma^2 I_N)^{-1} HC_\theta \]  
(6.17c)

\[ = C_\theta - (C_\theta H^T H + \sigma^2 I_{n_\theta})^{-1} C_\theta H^T H C_\theta \]  
(6.17d)

where (6.17b) and (6.17d) are proven in Appendix 6.6 on page 106. Note that (6.17b) and (6.17d) are time- and memory-efficient implementations of, respectively, (6.17a) and (6.17c).

From (6.17b) we conclude that

\[ \hat{\theta}_{\text{Bayes}}(y) = (C_\theta H^T H + \sigma^2 I_{n_\theta})^{-1} C_\theta H^T y \]  
(6.18)

Comparing the Bayes estimate (6.18) to the regularized linear least squares estimate (3.75) on page 54, it can be seen that both are equal if the following choices are made: \( P = C_\theta \) and \( \gamma = \sigma^2 \). If the covariance \( C_\theta \) increases to infinity, then (6.18) reduces to the linear least squares estimate (3.3a) on page 36.

For a fixed model, with uncertain prior knowledge (6.13), the expected value and covariance of \( \hat{\theta}_{\text{Bayes}}(y) \) (6.18) w.r.t. the measurements \( y \) equals

\[ E\{\hat{\theta}_{\text{Bayes}}(y)\} = (C_\theta H^T H + \sigma^2 I_{n_\theta})^{-1} C_\theta H^T H \theta_0 \]  
(6.19a)

\[ \text{Cov}\left(\hat{\theta}_{\text{Bayes}}(y)\right) = (C_\theta H^T H + \sigma^2 I_{n_\theta})^{-1} C_\theta H^T \sigma^2 I_N \]

\[ HC_\theta (H^T H C_\theta + \sigma^2 I_{n_\theta})^{-1} \]  
(6.19b)

Note that (6.17d) is different from (6.19b) because it quantifies the covariance of the model parameters, given the data, over all random realization of the model with Gaussian prior (6.13).

### 6.4. Gaussian Process Modeling

Gaussian process (GP) modeling is a special case of the Bayesian framework where the data \( z \) and the prior on the model parameters \( \theta \) is Gaussian (Rasmussen and Williams, 2006). Hence, for Gaussian process modeling, the three estimators (6.2) are equivalent and equal to the Bayes estimate (6.3b).

A difference with the classical Bayesian framework is that Gaussian process modeling handles the issue that in most practical cases only partial prior knowledge about the covariance (and the mean) of the model parameters is available [see Section 6.4.1].
The whole procedure is illustrated in Section 6.4.3 on the estimation of a finite impulse response model of a discrete-time system as elaborated in Chen et al. (2012) and Pillonetto et al. (2014).

6.4.1. Estimation Procedure

To solve the partial prior knowledge problem, the covariance (kernel) matrix of the model parameters is parametrized with some hyper-parameters $\psi$. The joint Gaussian pdf of the data $z$ and the model parameters $\theta$, necessary to construct the posterior pdf of $\theta$ (6.1b), is then conditioned on these hyper-parameters $\psi$

\[
 f_{z,\theta|\psi}(z, \theta|\psi)
\]  

Next, the marginal likelihood function $f_{z|\psi}(z|\psi)$ is constructed by marginalizing out the model parameters $\theta$ from the joint pdf (6.20)

\[
 f_{z|\psi}(z|\psi) = \int_{\Theta} f_{z,\theta|\psi}(z, \theta|\psi) d\theta
\]

where $\Theta$ is the domain of $\theta$. Next, $\psi$ is tuned by minimizing the negative log marginal likelihood

\[
 \hat{\psi} = \arg\min_{\psi} - \log f_{z|\psi}(z|\psi)
\]

Finally, using (6.14), the posterior mean (6.2a) is calculated for $\psi = \hat{\psi}$, giving

\[
 \hat{\theta}_{GP}(z) = \mathbb{E}\{\theta|z, \hat{\psi}\}
\]

Notes:

1. Since (part of) the prior parameter distribution is estimated from the data [see (6.22)], (6.23) is called the empirical Bayes estimate (Carlin and Louis, 2000).

2. For Gaussian distributions, the marginalization (6.21) and the posterior mean (6.23) calculations are very simple as illustrated in Sections 6.4.2 and 6.4.3.

6.4.2. Gaussian Process Linear Regression

Consider the linear regression problem of Section 6.3.2 and assume that the noise variance $\sigma^2$ is unknown and that the prior parameter covariance $C_{\theta}$ is parametrized

\[
 C_{\theta}(\alpha)
\]

101
with \( \alpha \in \mathbb{R}^{n_\alpha \times 1} \). The hyper-parameters \( \psi \) of the Gaussian process modeling problem are then

\[
\psi = \begin{bmatrix} \sigma^2 \\ \alpha \end{bmatrix}
\]  

(6.25)

Using (6.11), the mean and covariance of the data \( y \), given the hyper-parameters \( \psi \), are readily found

\[
\mathbb{E}\{y|\psi\} = 0 \\
\text{Cov}\{y|\psi\} = HC_\theta(\alpha)H^T + \sigma^2 I_N
\]  

(6.26a) (6.26b)

Hence, the negative logarithm of marginal likelihood function (6.21) equals

\[
-2 \log f_{y|\psi}(y|\psi) = y^T \left( HC_\theta(\alpha)H^T + \sigma^2 I_N \right)^{-1} y \\
+ \log \det \left( HC_\theta(\alpha)H^T + \sigma^2 I_N \right) + N \log(2\pi)
\]  

(6.27)

Minimizing (6.27) has shown to provide a trade-off between data-fit [first term in (6.27)] and model complexity [second term in (6.27)]. Other criteria for tuning the hyper-parameters do exist, for example, leave-one-out cross-validation (Rasmussen and Williams, 2006). Note that the cost function (6.27) is non-convex in the hyper-parameters \( \alpha \) and \( \sigma^2 \) and, hence, is prone to local minima. Finding good initial values for the hyper-parameters, to avoid local minima during the minimization of (6.27), is the bottle neck of the Gaussian process approach.

Finally, the minimizing argument \( \hat{\psi} \) of (6.27) is used to calculate the posterior mean (6.23)

\[
\hat{\theta}_{\text{GP}}(y) = (C_\theta(\hat{\alpha})H^TH + \hat{\sigma}^2 I_{n_y})^{-1}C_\theta(\hat{\alpha})H^Ty
\]  

(6.28)

which is (6.18) evaluated at \( \psi = \hat{\psi} \).

### 6.4.3. Estimation of a Finite Impulse Response Model

Consider the estimation of the impulse response coefficients \( g(t) \) of a discrete-time system from \( N \) known input \( u(t) \) and \( N \) noisy output \( y(t) \) samples, \( t = 0, 1, \ldots, N - 1 \)

\[
y(t) = \sum_{n=0}^{\infty} g(n)u(t-n) + v(t)
\]  

(6.29)

where \( v(t) \) is zero mean white Gaussian noise with variance \( \sigma^2 \). Since only a finite number of impulse response coefficients can be estimated from \( N \) input-output samples, the infinite convolution product in (6.29) is approximated by
an finite impulse response (FIR) model of order $R$

$$\sum_{r=0}^{R} g(r) u(t - r)$$  \hspace{1cm} (6.30)

Note that to calculate the first $R$ samples $t = 0, 1, \ldots, R - 1$, of the convolution product (6.30), one needs $R$ past values of the input, $u(-R), u(-R + 1), \ldots, u(-1)$, which are unknown. Therefore, the first $R$ output samples are discarded to estimate $g(r), r = 0, 1, \ldots, R$. This leads to a linear least squares problem (3.2) on page 36, with

$$y = \begin{bmatrix} y(R) & y(R + 1) & \ldots & y(N - 1) \end{bmatrix}^T$$  \hspace{1cm} (6.31a)

$$\theta = \begin{bmatrix} g(0) & g(1) & \ldots & g(R) \end{bmatrix}^T$$  \hspace{1cm} (6.31b)

$$H_{[k,l]} = u(R + k - l) \text{ with } \begin{cases} k = 1, 2, \ldots, N - R \\ l = 1, 2, \ldots, R + 1 \end{cases}$$  \hspace{1cm} (6.31c)

and where the regression matrix $H$ is Toeplitz (Ljung, 1999). From (6.31c), it follows that $N \geq 2R + 1$ samples are needed for estimating the FIR model (6.30) of order $R$.

In the Gaussian process modeling framework, the impulse response coefficients $\theta$ are regarded as a particular realization of a Gaussian process with zero mean value and a parametrized covariance matrix $C_\theta(\alpha)$. Among the different existing parametrizations, the diagonal and correlated covariance (kernel)

$$C_{\theta[k,l]} = \begin{cases} c \rho^{|k-l|} \lambda^{(k+l)/2} & k, l = 1, 2, \ldots \\ 0 & k, l = 0, -1, \ldots \end{cases}$$  \hspace{1cm} (6.32)

with hyper-parameters $c \geq 0, 0 < \lambda < 1$ and $|\rho| \leq 1$, has some interesting system properties [see Chen et al., 2012 and Pillonetto et al., 2014]:

1. Stability: the hyper-parameter $\lambda$ imposes the exponential decay of the impulse response.

2. Smoothness: the hyper-parameter $\rho$ quantifies the correlation between the impulse response coefficients $g(k - 1)$ and $g(l - 1)$. The larger the correlation, the smoother the impulse response and vice versa. Hence, $\rho$ tunes the complexity of the FIR model (6.30).

3. Causality: the impulse response coefficients $g(t)$ are zero for $t < 0$.

4. Prior knowledge versus data: the hyper-parameter $c$ makes a trade-off between the prior knowledge and the data.
Figure 6.2: Estimation of a finite impulse response model of order $R = 50$ from $N - R$ known input, noisy output samples. Left: first $R + 1$ samples of the true impulse response. Right: true impulse response [black] and the root mean squared error (RMSE) of the linear least squares [light and dark blue lines] and Gaussian process [coinciding light and dark red lines] estimates. Dark lines: $N - R = R + 1$. Light lines: $N - R = R + 2$.

The hyper-parameters $\psi$

$$\psi = \begin{bmatrix} \sigma^2 \\ \alpha \end{bmatrix} \quad \text{with} \quad \alpha = \begin{bmatrix} c \\ \rho \\ \lambda \end{bmatrix} \quad (6.33)$$

are tuned by minimizing the negative logarithm of marginal likelihood function (6.21). Using (6.26), we find (6.27), where $y$, $\theta$ and $H$ are defined in (6.31), and where $N$ is replaced everywhere by $N - R$. Finally, the Gaussian process estimate of the impulse response coefficients is obtained as in (6.28).

The performance of the Gaussian process modeling approach is illustrated on a second order discrete-time system with transfer function

$$G_0(z^{-1}) = \frac{0.58761z^{-1} + 0.53813z^{-2}}{1 - 0.65202z^{-1} + 0.77777z^{-2}} \quad (6.34)$$

It is the step-invariant (ZOH) transform of a second order continuous-time system

$$G(s) = \frac{\omega_0^2}{s^2 + 2\zeta\omega_0s + \omega_0^2} \quad (6.35)$$

with $\zeta = 0.1$, $f_0 = \omega_0/(2\pi) = 100$ Hz and $f_s = 500$ Hz. The left plot of Figure 6.2 shows the first 51 samples of the impulse response $g_0(n)$ of (6.34). Discrete-time system (6.34) is excited with zero mean white Gaussian noise with standard deviation one, and $N$ samples of the response are calculated.
These output samples are disturbed by additive, zero mean, white Gaussian noise, with standard deviation 0.1, resulting in a signal-to-noise-ratio of 24 dB. The FIR model (6.30) is estimated using (i) the linear least squares method calculated via the SVD of the scaled regression matrix [see (3.27a)], and (ii) the Gaussian process method (6.28) using the DC-kernel (6.32). The whole procedure is repeated for 400 independent realizations of the input and the disturbing noise.

The right plot of Figure 6.2 compares the root mean squared error (RMSE) of the estimates [blue and red lines] for two different values of $N$ to the true impulse response [black line]. For the case $N = R + 1$, the linear least squares estimate has a very high variability [dark blue line], while the RMSE of the Gaussian process estimate is almost 40 dB lower. Adding 1 data sample, $N = R + 2$, decreases significantly the RMSE of the linear least squares estimate, while the RMSE of the Gaussian process estimates remains almost unchanged. It can also be seen that no $[N = R + 1]$ or about 17 $[N = R + 2]$ impulse response coefficients can meaningfully be estimated using linear least squares, while in each case about 30 coefficients can be identified by the Gaussian process method.
6.5. Appendix: Proof of (6.15)

The proof is based on the factorization of a block matrix using its Schur complement. Consider the block matrix $M \in \mathbb{R}^{(m+n)\times(m+n)}$

$$M = \begin{bmatrix} A & C \\ B & D \end{bmatrix}$$

(6.36)

where $A \in \mathbb{R}^{n \times n}$, $B \in \mathbb{R}^{m \times n}$, $C \in \mathbb{R}^{n \times m}$ and $D \in \mathbb{R}^{m \times m}$. If the $D$ is regular ($\det(D) \neq 0$), then the matrix $M$ (6.36) can be factorized as

$$M = \begin{bmatrix} I_n & CD^{-1} \\ 0 & I_m \end{bmatrix} \begin{bmatrix} A - CD^{-1}B & 0 \\ 0 & D \end{bmatrix} \begin{bmatrix} I_n & 0 \\ D^{-1}B & I_m \end{bmatrix}$$

(6.37)

with $A - CD^{-1}B$ the Schur complement of $M$. The determinant and the inverse of $M$ (6.37) are given by

$$\det(M) = \det(A - CD^{-1}B)\det(D)$$

(6.38a)

$$M^{-1} = \begin{bmatrix} I_n & 0 \\ -D^{-1}B & I_m \end{bmatrix} \begin{bmatrix} (A - CD^{-1}B)^{-1} & 0 \\ 0 & D^{-1} \end{bmatrix} \begin{bmatrix} I_n & -CD^{-1} \\ 0 & I_m \end{bmatrix}$$

(6.38b)

[proof: it can easily be verified that $MM^{-1} = M^{-1}M = I_{n+m}$].

Consider now two jointly Gaussian distributed random vectors $x_1 \in \mathbb{R}^{m \times 1}$ and $x_2 \in \mathbb{R}^{n \times 1}$ with mean and covariance (6.14). Using (6.38), we readily find the conditional distribution of $x_1$, given $x_2$,

$$f_{x_1|x_2}(x_1|x_2) = \frac{f_{x_1,x_2}(x_1,x_2)}{f_{x_2}(x_2)} = \frac{1}{\sqrt{(2\pi)^n\det(S)}} e^{-\frac{1}{2}(z_1 - \Sigma_{12}\Sigma_{22}^{-1}z_2)^T S^{-1}(z_1 - \Sigma_{12}\Sigma_{22}^{-1}z_2)}$$

(6.39)

where $z_1 = x_1 - \mu_1$, $z_2 = x_2 - \mu_2$, and with $S = \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{12}^T$ the Schur complement of (6.14b). From (6.39) immediately follows (6.15), which concludes the proof.

6.6. Appendix: Proof of (6.17b) and (6.17d)

The proof is based on the following shift property of matrices. Consider two matrices $A \in \mathbb{R}^{n \times m}$ and $B \in \mathbb{R}^{m \times n}$. Then, for any $\lambda \neq 0$, it holds that

$$A(\lambda I_m + BA)^{-1} = (\lambda I_n + AB)^{-1}A$$

(6.40)
[proof: left and right multiplication with, respectively, $\lambda I_n + AB$ and $\lambda I_m + BA$, shows the equality]. Applying (6.40) to (6.17a) and (6.17c), with $A = C_0 H^T$ and $B = H$, proves (6.17b) and (6.17d), respectively.
Chapter 7
Neural Networks

Abstract: Neural network modeling has regained a lot of interest in the 21-st century because of (i) the increased computer memory and computation power, (ii) the improved training algorithms, and (iii) the world wide availability of “big data”. In this chapter we present the basic neural network topologies, and discuss the peculiarities of training (learning) deep neural networks for big data problems.

Learning Objectives:

- Basic neural network topologies and their constituting components.
- Training of deep neural networks: avoiding overfitting – early stopping, initialization of the weights and scaling of the input data.

7.1. Introduction

Neural networks, sometimes called artificial neural networks, are models inspired by the biological neural networks of animal brains. A neural network consists of a number of connected nodes, called artificial neurons, which are a simplified representation of neurons in biological brains [see Figure 7.1]. The output of each neuron is a nonlinear function of a weighted sum of the inputs plus some threshold value. The weights increase or decrease the strength of an input, and the threshold determines whether the output is active or not. Neurons are typically organized in multiple layers that perform different transformations on their inputs. The weights and threshold of each neuron constitute the parameters of the neural network model.

First, we define the neural network models commonly used [Section 7.2]. Next, we discuss the peculiarities of determining the weights and thresholds
from data [Section 7.3]. Finally, the theory is illustrated on the function approximation by feedforward neural networks [Section 7.4].

A comprehensive overview of deep neural network learning techniques and their use in big data, computer vision, speech recognition, natural language processing, and dynamical systems and control can be found in Goodfellow et al. (2016) and Brunton and Kutz (2019).

7.2. Neural Network Models

In this section we first define the basic components of a neural network [Section 7.2.1]. Next, using these components, some basic network topologies are presented [Section 7.2.2]. Finally, the universal approximation property of shallow and deep feedforward neural networks is discussed.

7.2.1. Components of a Neural Network

A neural network consists of three basic components: (i) input and output signal nodes [see Figure 7.2], (ii) artificial neurons [see Figure 7.3, top], and (iii) output neurons [see Figure 7.3, bottom]. The artificial neuron is a simplified representation of the biological neuron shown in Figure 7.1.

The output \( z \) of an artificial neuron is a nonlinear function \( \phi \) – called activation function – of a weighted sum of the inputs \( x_i, i = 1, 2, \ldots, n \), plus a threshold value \( b \)

\[
\begin{align*}
z &= \phi \left( b + \sum_{i=1}^{n} w_i x_i \right) \\
&= \phi \left( \sum_{i=0}^{n} w_i x_i \right)
\end{align*}
\]

(7.1a) (7.1b)

where \( w_i, i = 1, 2, \ldots, n \), are the weights. For notational convenience, the response \( z \) is often written under the form (7.1b), with \( x_0 = 1 \) and \( w_0 = b \).

Depending on the intended application of the neural network, different types of activation functions \( z = \phi(x) \) are used. For example, see Figure 7.4,

1. **Sigmoid functions**: these are monotonically increasing functions \( \phi(x) \) ranging between 0 \( (x = -\infty) \) and 1 \( (x = \infty) \). Examples are the logistic
Figure 7.1: Biological neuron and myelinated axon, with signal flow from inputs at the dendrites to outputs at the axon terminals [from https://en.wikipedia.org/wiki/Artificial_neural_network by Prof. Loc Vu-Quoc, University of Florida].

\[ u_n \rightarrow \phi \rightarrow y_m \]

Figure 7.2: Input (left) and output (right) signal nodes of a neural network.

\[ x_1 \]
\[ x_2 \]
\[ \vdots \]
\[ x_n \]
\[ z = \phi(b + \sum_{i=1}^{n} w_i x_i) \]
\[ y = b + \sum_{i=1}^{n} w_i x_i \]

Figure 7.3: Artificial neurons of a neural network. Top: artificial neuron with a nonlinear activation function $\phi$. It is a simplified version of the biologic neuron shown in Figure 7.1. Bottom: output neuron with an affine activation function.
and the squashing functions:

\begin{align*}
\text{logistic: } \phi(x) &= \frac{1}{1 + e^{-x}} & (7.2a) \\
\text{squashing: } \phi(x) &= \begin{cases} 0 & x < 0 \\ x & 0 \leq x \leq 1 \\ 1 & x > 1 \end{cases} & (7.2b)
\end{align*}

2. **Hyperbolic tangent function:**

\[ \phi(x) = \tanh(x) = \frac{e^x - e^{-x}}{e^x + e^{-x}} \]  

(7.2c)

3. **Softsign function:**

\[ \phi(x) = \frac{x}{1 + |x|} \]  

(7.2d)

4. **Rectified linear unit (ReLU):**

\[ \phi(x) = \begin{cases} 0 & x < 0 \\ x & x \geq 0 \end{cases} \]  

(7.2e)

5. **Radial basis functions:** these are functions that depend on the norm of the difference between the input vector \( x = [x_1 \, x_2 \ldots \, x_n]^T \) and a center \( c \in \mathbb{R}^{n \times 1} \). Usually, the Euclidean distance and a Gaussian function are taken:

\[ \phi(x) = e^{-\beta \|x - c\|^2_2} \]  

(7.2f)

where \( \beta > 0 \).

An exhaustive overview of activation functions and their properties can be found at

https://en.wikipedia.org/wiki/Activation_function

An output node (neuron) has typically a linear activation function [see Figure 7.3, bottom]. This is necessary for obtaining the universal function approximation property of a feedforward neural network [see Section 7.2.3].
Figure 7.4: Some commonly used activation functions $\phi(x)$: the logistic function (7.2a) [top left], the radial basis function (7.2f) with $\beta = 1$ [top right], the hyperbolic tangent function [middle left], the softsign function [middle right], the squashing function (7.2b) [bottom left], and the rectified linear unit (7.2e) [bottom right].
7.2.2. Neural Network Topologies

One can basically distinguish three main neural network topologies:

1. **Feedforward**: A feedforward neural network – formerly called *multilayer perceptron* – consists of an input layer, one or more hidden layers with nonlinear activation functions, and an output layer with linear activation functions [see Figure 7.5, top block diagram]. The width of each layer can be different, as well as the activation functions used. A *deep neural network* is a feedforward neural network with multiple hidden layers. Feedforward neural networks are used for interpolation, function approximation and classification.

2. **Recurrent**: A recurrent neural network is a feedforward neural network where the input layer and/or one or more hidden layers contain time-delayed versions of one or more outputs or downstream internal network signals. In the former case one speaks about input-output recurrent neural networks, while in the latter case of state space recurrent neural networks [see Figure 7.5, middle and bottom block diagrams]. Recurrent neural networks are used to model discrete-time nonlinear dynamical systems. Applications of recurrent neural networks include time series predictions (e.g. stock market), speech recognition and synthesis, text translation, grammar learning, handwriting recognition, search engine results, ...

3. **Convolutional**: A convolutional neural network is a feedforward neural network that contains one or more convolutional layers that act on tensor inputs. These networks are used for image and video processing.

A complete chart of neural network topologies can be found at

https://towardsdatascience.com/the-mostly-complete-chart-of-neural-networks-explained-3fb6f2367464

including a brief description of their use [see also Figure 7.6].

7.2.3. Universal Approximation Property

The function approximation properties of feedforward neural networks have been studied extensively in the literature for the two following cases: (i) fixed depth (number of layers) and increasing width (number of nodes in a layer), and (ii) fixed width and increasing depth.
Figure 7.5: Examples of feedforward (top), input-output recurrent (middle), and state space recurrent (bottom) neural networks. The internal nodes have a nonlinear activation function $\phi$, while the activation function of the output nodes is linear (oblique straight lines). $q$ delays the signal one time sample $qy(t) = y(t - 1)$. 

115
Figure 7.6: Neural network topologies [from https://towardsdatascience.com].
1. **Fixed depth, increasing width:**

   (a) Any function \( y = f(u) \) that is continuous over \( u \in [u_1, u_2] \), can be approximated arbitrarily well over \( u \in [u_1, u_2] \) by a feedforward neural network with one hidden layer and any sigmoidal (Hornik et al., 1989; Barron, 1993) or radial basis (Park and Sandberg, 1991, 1993) activation function, as the number of nodes \( n \) in the hidden layer increases to infinity.

   (b) If the function \( f(u) \) has continuous second order derivative, then the uniform approximation error is an \( O(n^{-1/2}) \) (Barron, 1993).

   (c) Some more general versions of the universal approximation theorem with relaxed conditions on the activation function can be found in Hornik (1991).

2. **Fixed width, increasing depth:**

   (a) Any function \( y = f(u) \) with continuous derivative over \( u \in [u_1, u_2] \), can be approximated arbitrary well over \( u \in [u_1, u_2] \) by a feedforward neural network, with fixed width \( n \) and ReLU activation functions \((7.2e)\), as the number of hidden layers increases to infinity (Yarotsky, 2017). Given the uniform approximation error \( \epsilon \in (0, 1) \), the number of hidden layers is at most an \( O(1 - \log(\epsilon)) \).

   (b) Any function \( y = f(u) \) that is continuous over \( u \in [u_1, u_2] \), can be approximated arbitrary well over \( u \in [u_1, u_2] \) by a feedforward neural network, with fixed width 4 and any nonaffine continuous activation function with at least at one point a non-zero continuous derivative, as the number of hidden layers increases to infinity (Kidger and Lyons, 2020).

Note that these results rely on the fact that the output layer uses linear activation functions. Multivariate versions of the universal approximation theorems can be found in Barron (1993), Yarotsky (2017) and Kidger and Lyons (2020). In Yarotsky (2017) it is shown that deep ReLU networks (more than one hidden layer) approximate more efficiently (less artificial neurons needed) smooth functions than shallow networks (e.g. one hidden layer). This motivates the use of deep ReLU neural networks.
7.3. Training/Learning a Neural Network

The weights (and hyper-parameters) of a feedforward neural network are obtained by minimizing a cost function. For example, assuming that \( N \) input \( u(k) \in \mathbb{R}^{n_u \times 1} \) and output \( y(k) \in \mathbb{R}^{n_y \times 1} \) samples are available, we get

\[
V(y, \theta) = \frac{1}{2} \sum_{k=1}^{N} E^T(k, \theta) E(k, \theta) \tag{7.3a}
\]

\[
E(k, \theta) = y(k) - \Phi_L(W_L\Phi_{L-1}(W_{L-1} \ldots \Phi_1(W_1u(k)))) \tag{7.3b}
\]

where \( Z_l = \Phi_l(V_l) \in \mathbb{R}^{n_l \times 1} \), with \( V_l = W_l Z_{l-1} \), \( Z_0 = u(k) \) and \( Z_L = \hat{y}(k) \), is the vector function of the \( n_l \) activation functions \( \phi_l \) of the \( l \)-th layer written under the form (7.1b); and \( W_l \in \mathbb{R}^{n_l \times n_{l-1}} \) is the weighting matrix of the \( l \)-th layer, with \( n_0 = n_u \), \( n_L = n_y \), and \( L \) the number of hidden layers plus the output layer (\( \Phi_L \) is a linear vector function). Finally, \( \theta \) is the vector of all weights

\[
\theta = [\text{vec}(W_L)^T \text{vec}(W_{L-1})^T \ldots \text{vec}(W_1)^T]^T \tag{7.3c}
\]

where \( \text{vec}(X) \) put the columns of \( X \) on top of each other. Minimizing (7.3) w.r.t. \( \theta \) is called training or learning the neural network.

First, by means of a glossary, we relate the concepts used in neural network modeling to those used in system identification and regression analysis [Section 7.3.1]. Next, it is shown how (7.3a) can be minimized time-efficiently for large neural networks and large data sets via back propagation [Section 7.3.2] and stochastic gradient descent [Section 7.3.3]. Finally, the issue of overtraining is handled in Section 7.3.4.

7.3.1. Glossary

While the system identification and regression analysis community adopted a classical statistical approach to the modeling problem, the neural network community developed its own terminology for traditional statistical concepts. Therefore, we provide a glossary of the terms commonly used (Sjöberg et al., 1995):

- to minimize a cost function = to train, to learn
- to validate = to generalize
- model structure = network
- model parameters = weights (and hyper-parameters)
estimation data = training data

validation data = generalization data

overmodeling, overfitting = overtraining, overlearning

undermodeling = oversmoothing

step size $1/\lambda^2$ in gradient descent (4.32) = learning rate

In the neural network and machine learning communities the following terms are commonly used:

- **batch size**: Amount of training data entries seen before updating the weights of the network via the stochastic gradient descent method [see Section 7.3.3 on page 121].

- **epochs**: Parameter defining the number of times the entire training (identification) data set is seen by the learning (minimization) algorithm [see Section 7.3.3 on page 121].

- **validation (generalization) data set**: The validation data set is a subset of the data used only to assess the performance of a given (network) model. When the network starts to overfit (overmodel) the data, the error on the validation data typically begins to rise.

- **test data set**: Data set that is independent of the training (identification) and validation (generalization) data. It is used to assess the quality of the final model. If the error on the test set reaches a minimum at a significantly different number of epochs than the validation set error, this might indicate a poor division of the data in training and validation sets. Therefore, it is useful to plot the test set error during the training process.

### 7.3.2. Back Propagation

The cost function (7.3a) is minimized w.r.t. the weights (7.3c) using the gradient descent method in Section 4.4.4 on page 71

$$\Delta\theta^{[p+1]} = -\frac{1}{\lambda^2} \frac{\partial V(y, \theta^{[p]})}{\partial \theta^{[p]}}$$

(7.4)
It requires the calculation of the derivative of the cost function w.r.t. each weight of the neural network

\[
\frac{\partial V(y, \theta)}{\partial W_{l[n,m]}} = \sum_{k=1}^{N} \left( \frac{\partial E(k, \theta)}{\partial W_{l[n,m]}} \right)^T E(k, \theta)
\]

(7.5)

where \(W_{l[n,m]}\) is the weight of the \(m\)-th input of the \(n\)-th node in the \(l\)-th layer.

Applying the chain rule of the derivative, we find,

\[
\frac{\partial E(k, \theta)}{\partial W_{l[n,m]}} = -\frac{\partial \Phi_L}{\partial V_L} W_L \frac{\partial \Phi_{L-1}}{\partial V_{L-1}} W_{L-1} \cdots \frac{\partial \Phi_{l+1}}{\partial V_{l+1}} W_{l+1} n \phi'(V_{l[n]}^T) Z_{l-1[m]}
\]

(7.6)

where \(V_l = W_l Z_{l-1}, \partial \Phi_l / \partial V_l, l = 1, 2, \ldots, L,\) are diagonal matrices, and with \(1_n\) a vector of zeros except at entry \(n\) where it is equal to one. Hence, one term in the sum (7.5) can be written as

\[
\left( \frac{\partial E(k, \theta)}{\partial W_{l[n,m]}} \right)^T E(k, \theta) = -\phi'(V_{l[n]}^T) Z_{l-1[m]} 1_n^T \delta_{l+1}(k)
\]

(7.7a)

\[
\delta_{l+1}(k) = W_{l+1}^T \frac{\partial \Phi_{l+1}}{\partial V_{l+1}} \delta_{l+2}(k)
\]

(7.7b)

\[\vdots\]

\[
\delta_{L-1}(k) = W_{L-1}^T \frac{\partial \Phi_{L-1}}{\partial V_{L-1}} \delta_{L}(k)
\]

(7.7c)

\[
\delta_{L}(k) = W_{L}^T \frac{\partial \Phi_{L}}{\partial V_{L}} E(k, \theta)
\]

(7.7d)

From (7.7) it follows that to calculate the derivative w.r.t. a weight of the \(l\)-th layer, we first calculate the derivative of the last layer \(L\) (7.7d), then the derivative of the second last layer \(L-1\), and so on till layer \(l+1\). This procedure – called back propagation – is time-efficient because:

1. The vectors \(\delta_i(k), i = L, L-1, \ldots, 2\) and \(k = 1, 2, \ldots, N\), only have to be calculated once.

2. Each vector \(\delta_i(k)\) in (7.7d)--(7.7b) is calculated via a matrix vector product [forward propagation requires matrix products, which takes much more time].

3. The vector \(\delta_{l+1}(k)\) is the same for \(n = 1, 2, \ldots, n_l\) and \(m = 1, 2, \ldots, n_{l-1}\).
7.3.3. Stochastic Gradient Descent

For large-scale problems (big data), the computational burden of the gradient descent method (4.32) combined with the back propagation algorithm (7.7) can still be too large. The stochastic gradient descent optimization replaces the actual gradient calculated from the entire data set, by an estimate obtained from a randomly selected subset of the data. Proceeding in this way, computational burden is traded for a slower convergence rate (Spall, 2003).

At every iteration step, the gradient (7.5) over the \(N\) data samples, is replaced by a randomly selected sub-set of \(1 \leq M \leq N\) samples

\[
\frac{\partial V(y, \theta)}{\partial W_{[n,m]}} \approx \sum_{k=1}^{N} \nu_k \left( \frac{\partial E(k, \theta)}{\partial W_{[n,m]}} \right)^T E(k, \theta) \tag{7.8}
\]

for \(n = 1, 2, \ldots, n_l\) and \(m = 1, 2, \ldots, n_{l-1}\), and where \(\nu_k, k = 1, 2, \ldots, N\), are independently distributed random variables satisfying \(E\{\nu_k\} = 1\). For example, the \(M\) randomly selected samples have the same weight \(\nu_k = N/M\), and \(\nu_k = 0\) for all other indices. \(M\) is then called the batch size. After \(N/M\) iterations all data is – on average – seen by the training (minimization) algorithm. This is called an epoch. Typically, the training, the validation and the test errors (costs functions) are plotted during training as a function of the number of epochs.

An accelerated version of the stochastic gradient method is obtained as (Nesterov, 1983, 2018)

\[
\psi[p+1] = \theta[p] + \beta(\theta[p] - \theta[p-1]) \tag{7.9a}
\]

\[
\theta[p+1] = \psi[p+1] - \frac{1}{\lambda^2} \frac{\partial V(y, \theta)}{\partial \theta} \bigg|_{\theta=\psi[p+1]} \tag{7.9b}
\]

where \(\beta \geq 0\) is called the momentum parameter, and where the gradient of the cost is calculated in each iteration step as in (7.8). Guidance for choosing the step size \(1/\lambda^2\) and the momentum parameter \(\beta\) is given in, respectively, Spall (2003) and Assran and Rabbat (2020).

Several other extensions and variants of the stochastic gradient descent algorithm have been proposed. A comparative overview can be found in Kingma and Ba (2015) and Dogo et al. (2018).

7.3.4. Early Stopping

Very often, the number of parameters (weights) of a neural network is (much) larger than the amount of training data. This results in overfitting (overtraining) if the (stochastic) gradient descent minimization of the cost function is
executed till a local minimum is reached. In each iteration step of the minimization procedure, a better fit of the training (identification) data set is obtained. Up to a certain point, it also improves the performance on an independent validation (generalization) data set. However, past this point, the improved fit on the training data set will come at the cost of an increased validation (generalization) error. To avoid this unwanted behavior, the cost on a validation data set is monitored during the estimation (training) of the parameters on the training data set. This procedure works well if the training and validation data sets are independent:

1. The ideal case is that independent training and validation data sets are available. If not, then the training data is split into independent or uncorrelated training and validation data sets in the portion two to one (two third for training and one third for validation). For example,

   (a) Stationary time series data: take the first 2/3 as training data and the last 1/3 for validation. Proceeding in this way, and in case of a colored noise disturbance, only the first samples of the validation set will be correlated with the last samples of the training set.

   (b) Function approximation: as validation data set one should take one third of the samples, that are uncorrelated with the other two third and that cover the same input range as the whole data set.

2. Train the weights (estimate the parameters) of the neural network on the training set by minimizing (7.3) and evaluate the cost (7.3) on the validation set during the iterations of the (stochastic) gradient descent algorithm (7.4) combined with (7.5) or (7.8). For example, every fifth iteration for gradient descent and every fifth epoch for stochastic gradient descent [one epoch includes a number of iterations such that all date have been seen].

3. Stop the training (the minimization of the cost function) as soon as the validation cost increases. By default Matlab stops the training procedure after six consecutive weight updates leading to a higher validation cost.

4. Use the weights obtained in the previous iteration step as final result.

7.3.5. Initialization of the Weights – Scaling of the Inputs

Appropriate initializing the weights of a deep neural network is of key importance for successful training and it has become a research field on its own
Glorot and Bengio, 2010; He et al., 2015). To avoid vanishing or exploding gradients of the cost function, the following initializations have been proposed for the weights $W_l \in \mathbb{R}^{n_l \times n_{l-1}}$ in the $l$-th layer:

1. *Hyperbolic tangent activation functions:* independent and uniformly distributed random variables (Glorot and Bengio, 2010)

   \[ W_{l[i,j]} \in \frac{\sqrt{6}}{\sqrt{n_l + n_{l-1}}} U(-1, 1) \quad (7.10) \]

   with $i = 1, 2, \ldots, n_l$, $j = 1, 2, \ldots, n_{l-1}$, and where $U(-1, 1)$ is the uniform distribution over the interval $[-1, 1]$.

2. *ReLU activation functions:* independent, zero mean normally distributed random variables

   \[ W_{l[i,j]} \in \frac{\sqrt{2}}{\sqrt{n_{l-1}}} N(0, 1) \quad (7.11) \]

   with $i = 1, 2, \ldots, n_l$, $j = 1, 2, \ldots, n_{l-1}$, and where $N(0, 1)$ is the zero mean Gaussian distribution with unit variance.

Note that the standard deviation of the uniformly distributed random variables (7.10) equals

\[ \frac{\sqrt{2}}{\sqrt{n_l + n_{l-1}}} \quad (7.12) \]

Hence, if $n_l = n_{l-1}$, then (7.12) is a factor $\sqrt{2}$ smaller than (7.11).

Given the random initializations of the weights (7.10) and (7.11), one should also normalize the inputs to avoid vanishing or exploding gradients in the first hidden layer (Sola and Sevilla, 1997). Indeed, the inputs have often different physical units resulting in numerical values of different orders of magnitude. An affine transformation is applied to the inputs, shifting their mean and scaling their amplitude. For example, the sample mean is subtracted and the result is divided by the sample standard deviation. This basic idea has been extended to all intermediate signals in a deep neural network resulting in an accelerated training (Ioffe and Szegedy, 2015).

### 7.3.6. Regularization

Following the same lines of the linear least squares method [see Section 3.4 on page 53], a penalty term for model complexity can be added to the cost function (7.3). The factor $\gamma$ in (3.74) is then called the regularization rate. Other penalty terms can be added for tuning the model complexity: see Section 8.3.2 on page 132 for the details.
7.4. Simulation Examples: Function Approximation by Feedforward Neural Networks

In this section we illustrate the universal approximation property of shallow and deep feedforward neural networks [Section 7.2.3], and the early stopping algorithm for avoiding overtraining [Section 7.3.4]. In both cases the arctan function is approximated over the interval \([0, 10]\) by neural networks with the ReLU activation function (7.2e).

7.4.1. Illustration of the Universal Approximation Property

The arctan function \(y(u) = \arctan(u)\) is evaluated at \(N = 400\) values

\[
u(k) = 10 \frac{k - 1}{N - 1} \quad \text{with} \quad k = 1, 2, \ldots, N
\]  

(7.13)

of which \(N/2\) are used for training and \(N/2\) for validation

\[
\text{training : } u(2k + 1) \quad \text{with} \quad k = 0, 1, \ldots, \frac{N}{2} - 1
\]  

(7.14a)

\[
\text{validation : } u(2k) \quad \text{with} \quad k = 1, 2, \ldots, \frac{N}{2}
\]  

(7.14b)

Two feedforward neural networks with ReLU activation functions (7.2e) are trained on the noiseless data:

1. Two hidden layer (deep) neural network with \(m\) nodes in each hidden layer. Using representation (7.1b) this network has in total \(m^2 + 4m + 1\) weights.

2. One hidden layer (shallow) neural network with \(n\) nodes in the hidden layer. Using representation (7.1b) this network has in total \(3n+1\) weights.

Taking \(m = 50\) for the deep network, we obtain \(n = 900\) for the shallow network with the same number (2701) of weights. Note that the number of weights (network parameters) is much larger than the number of training samples. Both networks are trained with the resilient backpropagation algorithm of Matlab.

Figure 7.7 shows the results. From the top right plot it can be seen that the approximation error of the shallow (one hidden layer, blue line) network is smaller than that of the deep (two hidden layers, red line) network. The mean squared error (MSE) performance on the training and validation data sets as a function of the number of epochs (one cycle through the full training data
set) is shown in the middle row. For both networks the MSE behavior on the training and validation data sets is similar.

Since the MSE on the validation data set decreases monotonically, no overfitting can be detected. Increasing the number of epoch will further diminish the approximation errors of both networks, which illustrates their universal approximation property. However, it can be seen that the decrease rate of the deep network is smaller than that of the shallow network. This was confirmed by many independent training runs with random initialization of the weights. Each and every time the approximation error of the deep network is larger than that of the shallow network and it convergence to zero is slower. This is confirmed by a Monte Carlo simulation of hundred runs, which shows that the root mean squared error of the deep network is larger than that of the shallow network for $u \in [0, 5]$ [see the bottom plot of Figure 7.7]. What differs in each Monte Carlo run is the random initialization of the network weights.

This example illustrates that deep neural networks are harder to train than shallow networks. It also explains why the initialization and training of deep neural networks is still an active research topic in machine learning.

### 7.4.2. Illustration of the Early Stopping Algorithm

To illustrate the overfitting problem, the arctan function $y_0(u) = \arctan(u)$ is evaluated at $N = 20$ values

$$u(k) = 10 \frac{k - 1}{N - 1} \quad \text{with} \quad k = 1, 2, \ldots, N$$

(7.15)

Zero mean Gaussian noise $v(k)$ with standard deviation 0.25 is added to the function values

$$y(k) = y_0(k) + v(k)$$

(7.16)

for $k = 1, 2, \ldots, N$. Two independent data sets (7.16) of $N = 20$ samples each are generated, one for training and one for validation. These data sets are used for training a shallow (one hidden layer) feedforward neural network with $n = 900$ nodes. Clearly, the number of weights (network parameters) to be estimated (2701) is much larger than the amount of training samples (20). The whole procedure is repeated for 400 independent random realizations of the disturbing noise and the initialization of the network weights.

The top and middle rows of Figure 7.8 show the results for one realization of the 400 Monte Carlo runs. From the middle row it can be seen that the MSE on the training set [blue] decreases to zero as the number of epochs increases to infinity, while that of the validation data set remains unchanged.
Figure 7.7: Approximation of the arctan function [top left] over the interval [0, 10] by feedforward neural networks with ReLU activation functions (7.2e). Top right: approximation error $\hat{y} - y_0$ of the shallow [blue] and deep [red] neural networks with the same number of weights. Middle row: mean squared error (MSE) on the training [blue] and validation [red] data sets as a function of the number of epochs. Middle left: shallow network. Middle right: deep network. Bottom: ± root mean squared error (RMSE) [solid lines] and error mean value [dashed lines] over hundred Monte Carlo runs of the shallow [blue] and deep [red] neural networks.
after about 40 epochs. It illustrates the overtraining (overfitting) problem of neural networks: after more than 6000 epoch the neural network fits almost exactly the \( N = 20 \) noisy arctan values. This problem is avoided by selecting the neural network ['×'] corresponding to the minimum ['o'] of the validation MSE [middle right plot]. Note that for each Monte Carlo run this minimum will appear at a different epoch.

No conclusions concerning the variability of the early stopping algorithm can be drawn from one realization [top right plot]. Therefore, the root mean squared error (RMSE) over the 400 Monte Carlo runs of the two neural network estimates are compared in the bottom plot of Figure 7.8. It can be seen that the RMSE with early stopping [red] is indeed lower than without early stopping [blue]. Note that the RMSE without early stopping coincides with the noise standard deviation 0.25 [Explain why!].

Finally, from the bottom plot of Figure 7.8 it can also be seen that the error of the mean estimate with early stopping ['×'] almost falls within its 95% uncertainty bound [two times the RMSE value divided by the square root of 400]. It indicates that the early stopping algorithm makes a good bias-variance trade-off.
Figure 7.8: Approximation of the arctan function over the interval [0, 10] by a shallow (one hidden layer) feedforward neural network with ReLU activation functions (7.2e). Top left: true value [black], noisy data [black ×], and neural network estimate without [blue] and with early stopping [red]. Top right: approximation error \( \hat{y} - y_0 \) of the neural network estimates. Middle left: mean squared error (MSE) performance on the training [blue] and validation [red] data sets. Middle right: zoom of the middle left plot. The black ‘o’ and ‘×’ indicate the neural network selected by the early stopping algorithm. Bottom: root mean squared error (RMSE) over 400 Monte Carlo runs of the neural network estimates without [blue solid line] and with [red solid line] early stopping, and the error of the mean estimates with early stopping [black ‘×’] and its 95% uncertainty bound [red dashed lines].
Chapter 8
Tuning the Model Complexity

Abstract: Model selection is a critical issue in data based modeling: the model should not be too simple (bias errors) nor too complex (variance errors). In this chapter we present techniques to make an optimal trade-off between the bias and variance errors. Most of these techniques adjust the model complexity over a discrete set of integer numbers, while some others tune the model complexity in a continuous fashion.

Learning Objectives:

• The role of validation data in model selection.
• The importance of penalty terms for model selection.
• Techniques for discrete adjustment of the model complexity.
• Methods for continuous tuning of the model complexity

8.1. Problem Statement

A critical issue in data based modeling is the tuning of the model complexity. If the model is too simple, then the bias error is dominant, while if the model is too complex, then the variance error is most important. Making an appropriate bias-variance trade-off such that the mean squared error is minimal is the ultimate goal of the model selection algorithms.

The model selection algorithms can be divided in two main classes. The first class uses an independent validation data set to monitor the mean squared error as a function of the model complexity [see Section 8.2]. This procedure is similar to the early stopping algorithm used for training neural networks [see Sections 7.4.2 and 7.4.2 on, respectively, pages 121 and 125]. The second class adds a penalty term to the cost function to be minimized [see Section 8.3]. This penalty term becomes larger for increasing model complexity such that
the overall cost mimics the behavior of the mean squared error on a validation data set.

In the remainder of this chapter we first describe the main algorithms of both classes [Sections 8.2 and 8.3] and, next, illustrate their performance on the polynomial approximation of the arctan function [Section 8.4].

8.2. Use of Validation Data

The ideal case is that two (large) independent (uncorrelated) data sets are available: one for identification (training) and one for validation (generalization). If not, then the available data is split into a identification and validation set. Several approaches exist:

1. Split the data into an independent (uncorrelated) identification (training) and validation data set in the ratio two to one (2/3 for identification and 1/3 for validation). This approach is typically used in time series analysis and system identification where the first 2/3 samples are used for identification and the last 1/3 samples for validation. Note that in case of colored noise the last samples of the identification set are correlated with the first samples of the validation set.

2. Leave-one-out-cross-validation (LOOCV): All $N$ data samples, expect one independent/uncorrelated of the rest, is used for identification. This is repeated for all possible validation samples. This gives $N$ validation residuals to calculate a sum of squared residuals cost. The $N$ estimates can be averaged. The LOOCV approach is suitable for short data sets.

3. $p$-fold cross-validation: The $N$ data samples are randomly partitioned in $p$ independent sets of $N/p$ samples. One set of $N/p$ samples is retained for validation, the $p - 1$ other sets for training (identification). This is repeated $p$ times giving an MSE of the $p$ validation sets. The $p$ estimates can be averaged. Typically, a 10-fold cross-validation is used. The $p$-fold cross-validation method is appropriate for long data records.

In all cases the MSE on the validation set is calculated and monitored as a function of the model complexity. The model with the smallest validation MSE is selected. In a final step, one can use the selected model complexity to estimate the model parameters using all data (training plus validation data).

Note that the LOO and $p$-fold cross-validation algorithms use each sample exactly once for validation. Note also that it is important that the validation
data is independent of, or at least uncorrelated with, the training data. This should be kept in mind when the data is disturbed by colored (correlated) noise.

8.3. Use of Penalty Terms

Within this class of methods using a penalty term, one can distinguish two sub-classes: (i) the algorithms that tune the model complexity over a discrete set of integer numbers [Section 8.3.1], and (ii) the methods that tune the model complexity in a continuous manner [Section 8.3.2].

8.3.1. Discrete Tuning of the Model Complexity

The model selection criteria minimize a cost function of the form

$$V(\hat{\theta}(z), z) p(n_\theta, N)$$

over the considered discrete set of models, where $\hat{\theta}(z)$ is the minimizer of the sum of the squared residuals cost function $V(\theta, z)$, and with $p(n_\theta, z)$ a penalty term that increases with the model complexity (number of model parameters $n_\theta$). Different penalty terms have been constructed in the literature. The most important ones are the final prediction error (FPE) of Akaike (1969), Akaike's information criterion (AIC) described in Akaike (1974), and the minimum description length (MDL) – sometimes called Bayesian information criterion (BIC) – of Rissanen (1978) and Schwartz (1978)

- **FPE**: $p(n_\theta, N) = \frac{N + n_\theta}{N - n_\theta}$
- **AIC**: $p(n_\theta, N) = e^{2n_\theta} N$
- **MDL, BIC**: $p(n_\theta, N) = e^{n_\theta \log N} N$

with $n_\theta = \text{dim}(\theta)$ and $N$ the number of data samples. In the remainder of this section we give a motivation for criteria (8.1b) and (8.1c).

The final prediction error criterion adds a penalty term $2n_\theta \sigma^2$ to cost function (Akaike, 1969; Ljung, 1999)

$$V(\hat{\theta}(z), z) + 2n_\theta \sigma^2$$

where $\sigma^2$ is the disturbing noise variance. It follows from (8.2) that adding a parameter only pays off if the decrease of the cost function $V(\hat{\theta}(z), z)$ is larger
than $2\sigma^2$. Since $\sigma^2$ is mostly unknown, it should be estimated from the data. A suitable estimate is obtained as

$$\hat{\sigma}^2 = \frac{V(\hat{\theta}(z), z)}{N - n_\theta} \quad (8.3)$$

where $N - n_\theta$ are the degrees of freedom of the residuals. Substituting (8.3) in (8.2) gives (8.1b).

Assuming that the noise is normally distributed, a penalty $n_\theta$ is added to the negative log-likelihood function (Akaike, 1974; Ljung, 1999)

$$\frac{V(\hat{\theta}(z), z)}{2\sigma^2} + \frac{N}{2} \log \sigma^2 + \frac{N}{2} \log(2\pi) + n_\theta \quad (8.4)$$

Note that dividing (8.2) by $2\sigma^2$ results in the same additive penalty term as in (8.4). Minimizing (8.4) w.r.t. $\sigma^2$ gives an estimate of the unknown noise variance $\hat{\sigma}^2$

$$\hat{\sigma}^2 = \frac{V(\hat{\theta}(z), z)}{N} \quad (8.5)$$

Substituting (8.5) in (8.4), we find

$$\frac{N}{2} + \frac{N}{2} \log V(\hat{\theta}(z), z) - \frac{N}{2} \log N + \frac{N}{2} \log \sigma^2 + \frac{N}{2} \log(2\pi) + n_\theta \quad (8.6)$$

where only the second and the last term depend on the model complexity. Multiplying the sum of these two terms by $2/N$ and taking the exponential function, finally, gives (8.1c).

For small sample sizes ($n_\theta \sim N$) the AIC (8.1c) and MDL (8.1d) penalty terms are replaced by

$$\text{AIC : } p(n_\theta, N) = e^{\frac{2n_\theta}{N - n_\theta}} \quad (8.7a)$$

$$\text{MDL : } p(n_\theta, N) = e^{n_\theta \log \frac{N}{N - n_\theta}} \quad (8.7b)$$

(McQuarrie et al., 1997; De Ridder et al., 2005).

8.3.2. Continuous Tuning of the Model Complexity

Following the same lines of the regularized linear least squares method (3.74) on page 54, a penalty term can be added to the sum of squared residuals cost $V(\theta, z)$, resulting in a generalized $L_2$-regularization

$$V(\theta, z) + \gamma \theta^T P^{-1} \theta \quad (8.8)$$
where $\gamma \geq 0$ and $P$ is a (semi-)positive definite matrix. Minimizing (8.8) w.r.t. $\theta$ gives an estimate $\hat{\theta}(z)$ whose complexity is continuously tuned by the parameter $\gamma$ and the matrix $P$. For example, in Gaussian process modeling, the complexity is continuously tuned via $\gamma$ and the hyper-parameters $\alpha$ describing the matrix $P(\alpha)$ [see Section 6.4 on page 100]. The mean squared error is minimal for the choice $\gamma = \sigma^2$ and $P = \theta_0\theta_0^T$ [see Section 3.4 on page 53].

Taking $P = I_n$, one gets the classical $L_2$ regularization

$$V(\theta, z) + \gamma \|\theta\|_2^2 \quad (8.9)$$

Choosing $\gamma = \sigma^2$ gives the Tikhonov regularization (Tikhonov and Arsenin, 1977), that minimizes the mean squared error [see Section 3.4 on page 53].

Replacing $\|\theta\|_2^2$ by $\|\theta\|_1 = \sum_{i=1}^{n_\theta} |\theta_i|$ gives the $L_1$ regularization

$$V(\theta, z) + \gamma \|\theta\|_1 \quad (8.10)$$

also called LASSO (least absolute shrinkage and selection operator) or basis pursuit. Minimizing (8.10) w.r.t. $\theta$ selects a sparse solution (only a few entries of $\hat{\theta}(z)$ are non-zero), which is necessary in, for example, feature selection.

Combining (8.9) and (8.10) results in the elastic net regularization

$$V(\theta, z) + \gamma \left( \rho \|\theta\|_1 + (1 - \rho) \|\theta\|_2^2 \right) \quad (8.11)$$

where $\rho \in [0, 1]$. It avoids the possible non-uniqueness of the $L_1$ solution (8.10), at the cost of losing sparsity.

Note that (8.9) and (8.10) implicitly assume that the entries of $\theta$ have the same physical units, while (8.11) only makes sense if the model parameters have no physical units. Therefore, to apply the $L_2$ (8.9), $L_1$ (8.10), and elastic net (8.11) regularizations, one should first scale the model such that $\theta$ has no units.

Figure 8.1 compares the different solutions (8.8)–(8.11) for the two dimensional case assuming that the cost function $V(\theta, z)$ does not change along the straight line $\theta_2 = -\beta_1\theta_1 + \beta_0$, with $\beta_0, \beta_1 > 0$ [red line]. The intersection between the red straight line and the constant regularization term [black line] is the minimizer of the regularized cost functions (8.8)–(8.11). It can be seen that the $L_1$ and elastic net regularizations select a sparse solution $\theta_1 = 0$, which is not the case for the (generalized) $L_2$ regularizations.
8.4. Simulation Example: Polynomial Approximation of a Function

In this section we compare the performance of the LOOCV [Section 8.2] and AIC, MDL [Section 8.3.1] model selection criteria to that of the optimal $L_2$- and generalized $L_2$-regularizations [Section 8.3.2]. As an example we take the polynomial approximation of the arctan function over the interval $[-2, 4]$, under exactly the same simulation conditions as described in Section 3.4 on page 55. For each of the thousand Monte-Carlo runs, we calculate the following estimates, starting from $N = 17$ samples disturbed by zero mean Gaussian noise with standard deviation $\sigma_y = 0.5$:

1. The linear least squares estimate (3.28b) on page 40, for the fixed polynomial order $n_\theta - 1 = 14$.

2. The linear least squares estimate (3.28b) that minimizes the AIC criterion (8.1a), (8.7a) over $n_\theta - 1 \in [0, 14]$.

3. The linear least squares estimate (3.28b) that minimizes the MDL criterion (8.1a), (8.7b) over $n_\theta - 1 \in [0, 14]$. 

Figure 8.1: Constant value of the additive penalty (regularization) terms used for tuning the model complexity – two dimensional case. Top left: generalized $L_2$-regularization (8.8). Top right: $L_2$-regularization (8.9). Bottom left: $L_1$-regularization (8.10). Bottom right: elastic net regularization (8.11).
Figure 8.2: Model selection criteria [left] and its zoom [right] as a function of the polynomial order \( n_\theta - 1 \in [0, 14] \) for one realization of the thousand Monte-Carlo runs. Magenta: MDL model selection (8.1a) and (8.7b). Cyan: AIC model selection (8.1a) and (8.7a). Dark green: LOOCV model selection. For this realization, the three criteria select the polynomial order \( n_\theta - 1 = 2 \).

4. The linear least squares estimate (3.28b) that minimizes the LOOCV cost over \( n_\theta - 1 \in [0, 14] \).

5. The \( L_2 \)-regularized estimate (8.9), with \( \gamma = \sigma^2_y \) (= Tikhonov regularization).

6. The optimal generalized \( L_2 \)-regularized estimate (8.8), with \( \gamma = \sigma^2_y \) and \( P = \theta_0 \theta_0^T \), and where \( \theta_0 \) is obtained from a linear least squares fit on the noiseless data.

Figure 8.2 shows the AIC, MDL and LOOCV cost functions as a function of the model complexity for one particular Monte-Carlo run. It can be seen that for this particular realization, the cost functions all reach their minimum at \( n_\theta - 1 = 2 \) (= second order polynomial).

The polynomial orders \( n_\theta - 1 \) selected by AIC, MDL and LOOCV over the thousand Monte-Carlo runs are given in Table 8.1. It can be seen that the model selection is a random process with a criterion-dependent distribution. On average, the model complexity selected by MDL is smaller than that of AIC, which on its turn is smaller than that of LOOCV. Note that none of the criteria select a too simple \([n_\theta - 1 = 0]\) nor a too complex \([n_\theta - 1 > 9]\) model.

Figure 8.3 compares the performance (RMS, variance, and bias errors) of the six estimates. The following observations can be made:

a. The optimal generalized \( L_2 \) regularization [blue] has the smallest RMS, variance and bias errors.

135
Table 8.1: Number of times a polynomial order $n_\theta - 1 \in [0, 14]$ is selected by AIC, MDL and LOOCV over thousand Monte-Carlo runs.

<table>
<thead>
<tr>
<th>$n_\theta - 1$</th>
<th>AIC</th>
<th>MDL</th>
<th>LOOCV</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>1</td>
<td>326</td>
<td>494</td>
<td>249</td>
</tr>
<tr>
<td>2</td>
<td>410</td>
<td>380</td>
<td>324</td>
</tr>
<tr>
<td>3</td>
<td>117</td>
<td>75</td>
<td>157</td>
</tr>
<tr>
<td>4</td>
<td>91</td>
<td>45</td>
<td>141</td>
</tr>
<tr>
<td>5</td>
<td>30</td>
<td>4</td>
<td>62</td>
</tr>
<tr>
<td>6</td>
<td>20</td>
<td>2</td>
<td>36</td>
</tr>
<tr>
<td>7</td>
<td>4</td>
<td>0</td>
<td>27</td>
</tr>
<tr>
<td>8</td>
<td>2</td>
<td>0</td>
<td>3</td>
</tr>
<tr>
<td>9</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>10</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>11</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>12</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>13</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>14</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

b. The linear least squares estimate with the fixed polynomial order $n_\theta - 1 = 14$ [yellow] is unbiased [explain this!] but has the largest RMS and variance errors. The latter is equal to the disturbing noise variance [explain this!].

c. While the Tikhonov regularization [red lines] has the largest bias error, its variance error is the second smallest.

d. The AIC [cyan], MDL [magenta] and LOOCV [green] estimates perform more or less equally well [MDL has the highest bias but the smallest variance], and have RMS and bias errors that are smaller than those of the Tikhonov regularization.

e. For the Tikhonov regularization the bias and variance errors are equally important, while for the other estimators the variance error dominates.
Figure 8.3: Polynomial approximation of the arctan function – Results thousand Monte-Carlo runs. Dark yellow: polynomial of order $n_\theta - 1 = 14$. Dark red: Tikhonov regularization (8.9) with $\gamma = \sigma_y^2 = (0.5)^2$. Magenta: MDL model selection (8.1a) and (8.7b). Cyan: AIC model selection (8.1a) and (8.7a). Dark green: LOOCV model selection. Dark blue: optimal regularization (3.78). Top left: RMS error over the thousand runs. Top right: zoom of the RMS error. Bottom left: sample variance over the thousand runs. Bottom right: bias error of the mean value over the thousand runs.
Chapter 9
Evaluation Matrix

The emphasis lies on the understanding and the application of the concepts, not
the reproduction. It makes no sense to learn by heart the technical conditions
and assumptions leading to a particular theoretical result. However, one should
understand why they are made and where they are used in the proofs.

Mastering a data based modeling course goes through hands on experience.
Therefore, we strongly encourage the students

1. To make all on-paper exercises mentioned throughout the course.

2. To program in Matlab the three toy examples used throughout the whole
course: (i) the estimation of the resistor value from noisy DC measure-
ments, (ii) the estimation of the slope of a straight line through the
origin from noisy abscissa and ordinate values, and (iii) the polynomial
and neural network approximation of the arctan function (noiseless and
noisy cases).

Make a critical analysis of the results obtained, and based on this, draw some
relevant conclusions.

Evaluation Matrix.

<table>
<thead>
<tr>
<th>Learning Content</th>
<th>Knowledge</th>
<th>Insight</th>
<th>Apply</th>
</tr>
</thead>
<tbody>
<tr>
<td>Chapter 1</td>
<td>×</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Chapter 2</td>
<td>×</td>
<td>×</td>
<td>×</td>
</tr>
<tr>
<td>Chapter 3</td>
<td>×</td>
<td>×</td>
<td>×</td>
</tr>
<tr>
<td>Chapter 4</td>
<td>×</td>
<td>×</td>
<td>×</td>
</tr>
<tr>
<td>Chapter 5</td>
<td>×</td>
<td>×</td>
<td>×</td>
</tr>
<tr>
<td>Chapter 6</td>
<td>×</td>
<td>×</td>
<td>×</td>
</tr>
<tr>
<td>Chapter 7</td>
<td>×</td>
<td>×</td>
<td>×</td>
</tr>
<tr>
<td>Chapter 8</td>
<td>×</td>
<td>×</td>
<td>×</td>
</tr>
</tbody>
</table>
Bibliography


Nesterov, Y. *Lectures on Convex Optimization*. Springer, Cham (Switzerland), 2018.


144


Subject Index

activation functions
  definition, 110
  hyperbolic tangent, 112
  radial basis function, 112
  rectified linear unit, 112
  sigmoid functions, 110
  soft sign, 112
asymptotic covariance
  definition, 16
  function of a consistent estimator, 31
  via cost function, 29
asymptotic efficiency
  definition, 20
  via cost function, 31
asymptotic normality
  definition, 17
  via cost function, 31
asymptotically unbiased, 14
back propagation
  definition, 119
batch size
  definition, 121
Bayesian estimator
  asymptotic properties, 96
  connection with regularized linear least squares, 99
  definition, 95
  examples, 97
  Gaussian process modeling, 100
  linear regression, 99
bias compensated least squares
  definition, 46
  resistor example, 47
central limit theorem, 16
consistency
  definition, 14
  via cost function, 29
convergence in distribution, 13
convergence in law, 13
convergence in mean square, 13
convergence in probability, 13
convergence with probability one, 12
Cramér-Rao lower bound
  definition, 19
  linear least squares, 37
  nonlinear least squares, 66
  resistor measurement, 26
  sample mean, sample variance, 87
curve fitting example
  condition number, 41
  definition, 6
  neural network, 124
  outliers, 58
  rational approximation, 75
  regularization, 55
  tuning model complexity, 134
early stopping
  definition, 121
  illustration, 125
efficiency
definition, 20
epoch
definition, 121
examples
Bayesian linear regression, 99
curve fitting, 6, 41, 55, 58, 75,
124, 134
estimation of a finite impulse
response model, 102
exponential function, 74
resistor measurement, 3, 21,
47, 51, 87
sample mean, sample variance,
84
slope straight line, 8, 88, 97
sum harmonically related
sinewaves, 68

Fisher information matrix
definition, 19
linear least squares, 37
nonlinear least squares, 66
resistor measurement, 26
slope straight line, 91

Gauss-Newton, 70
Gaussian process modeling
connection with regularized
linear least squares, 99
definition, 100
estimation of a finite impulse
response model, 102
estimation procedure, 101
linear regression, 101
generalization data
definition, 130
generalized singular value
decomposition, 49
generalized total least squares
definition, 48
resistor example, 49
gradient descent, 71

identification data
definition, 130
instrumental variables
definition, 50
resistor example, 51

law of large numbers, 15
learning, see minimization cost
function
leave-one-out-cross-validation
definition, 130
Levenberg-Marquardt, 70
line search, 71
linear least squares
asymptotic normality, 38
consistency, 37
definition, 36
formula, 36
inefficiency, 37
numerical stable calculation,
38
optimal regularization, 54
outliers, 56
regularization, 53
Tikhonov regularization, 54
unbiasedness, 36
weighting, 44
linear least squares – noisy
regression matrix
bias compensated, 46
definition, 44
instrumental variables, 50
known regression matrix, 46
observed regression matrix, 45
maximum likelihood
asymptotic covariance, 83
asymptotic efficiency, 83
consistency, 81
definition, 79
examples, 84
robustness, 84
minimization cost function
algorithms, 68
early stopping, 121
Gauss-Newton, 70
gradient descent, 71
Levenberg-Marquardt, 70
line search, 71
Newton-Raphson, 69
starting values, 73
stochastic gradient descent, 121
stopping criteria, 72
model complexity
illustration, 134
leave-one-out-cross-validation, 130
p-fold cross-validation, 130
penalty term, 131
regularization, 133
validation data, 130
neural network
activation function, 110
back propagation, 119
batch size, 121
convolutional, 114
definition, 110
early stopping, 121
epoch, 121
feedforward, 114
initialization weights, 122
recurrent, 114
regularization, 123
stochastic gradient descent, 121
training/learning, 118
universal approximation property, 114
Newton-Raphson, 69
nonlinear least squares
asymptotic covariance, 66
asymptotic normality, 66
consistent, 65
definition, 63
minimization cost function, 68
regularization, 132
separable least squares, 67
variable projection, 67
outliers
curve fitting example, 58
definition, 56
p-fold cross-validation
definition, 130
penalty term
Akaike information criterion, 131
definition, 131
final prediction error criterion, 131
minimum description length, 131
short data size, 132
regularization
$L_1$, 133
$L_2$, 133
curve fitting example, 55
elastic net, 133
linear least squares, 53
nonlinear least squares, 132
Author Index

Abel (1993), 20
Akaike (1969), 131
Akaike (1974), 131, 132
Assran and Rabbat (2020), 121
Barron (1993), 117
Ben-Israel and Greville (1974), 39, 54
Billingsley (1995), 17
Boyd and Vandenberghe (2004), 58, 73
Brillinger (1981), 17, 18
Brunton and Kutz (2019), 110
Caines (1988), 80
Carlin and Louis (2000), 101
Chen et al. (2012), 54, 101, 103
Chow and Teicher (1988), 17
De Ridder et al. (2005), 132
Dogo et al. (2018), 121
Feller (1968), 17
Fletcher (1991), 68, 70, 71
Glorot and Bengio (2010), 123
Golub and Pereyra (1973), 67
Golub and Van Loan (1996), 38, 39, 44, 49
Goodfellow et al. (2016), 110
Hampel et al. (2005), 58
He et al. (2015), 123
Hornik (1991), 117
Hornik et al. (1989), 117
Huber (1981), 57
Ioffe and Szegedy (2015), 123
Jazwinski (1970), 13
Kaplan (1993), 30, 32
Kidger and Lyons (2020), 117
Kingma and Ba (2015), 121
Ljung (1999), 103, 131, 132
Lukacs (1975), 12, 13, 15
Markovsky and Van Huffel (2007), 48, 76
McQuarrie et al. (1997), 132
Nelles (2001), 73
Nesterov (1983), 121
Nesterov (2018), 121
Norton (1986), 51
Park and Sandberg (1991), 117
Park and Sandberg (1993), 117
Pillonetto et al. (2014), 54, 101, 103
Pintelon and Hong (2007), 90
Pintelon and Schoukens (1996), 68
Pintelon and Schoukens (2012), 16, 17, 38, 41, 49, 64–66, 76
Pintelon et al. (1998), 49
Press et al. (2007), 73
Ralston and Rabinowitz (1984), 69
Rasmussen and Williams (2006), 100, 102
Rissanen (1978), 131
Schwartz (1978), 131
Sjöberg et al. (1995), 118
Söderström and Stoica (1989), 53
Söderström (1974), 29
Söderström (2013), 51
Sola and Sevilla (1997), 123
<table>
<thead>
<tr>
<th>Reference</th>
<th>Page Numbers</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sorenson (1980)</td>
<td>80</td>
</tr>
<tr>
<td>Spall (2003)</td>
<td>121</td>
</tr>
<tr>
<td>Stout (1974)</td>
<td>12, 16</td>
</tr>
<tr>
<td>Stuart and Ord (1994)</td>
<td>96</td>
</tr>
<tr>
<td>Tikhonov and Arsenin (1977)</td>
<td>54, 133</td>
</tr>
<tr>
<td>Van Barel and Bultheel (1992)</td>
<td>41</td>
</tr>
<tr>
<td>Van Huffel and Vandewalle (1991)</td>
<td>48, 49, 76</td>
</tr>
<tr>
<td>Yarotsky (2017)</td>
<td>117</td>
</tr>
</tbody>
</table>