Universität Bremen

Zentrum für Technomathematik

# Diplomarbeit

zur Erlangung des akademischen Grades
Diplom-Technomathematiker (Dipl.-Math.techn.)

Universität Bremen
Zentrum für Technomathematik

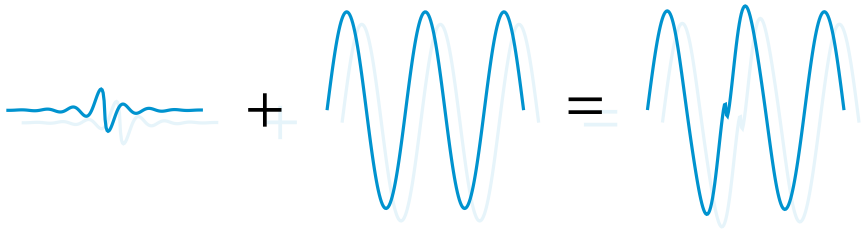
---

# Sparsity in Geosciences

## Sparse Decomposition for Analysis of Sea Floor Pressure Data

---

Matthias Ehrhardt



vorgelegt am 21.11.2011 von

Matthias Ehrhardt
Matrikel-Nr.: 2130018
ehrhardt@math.uni-bremen.de

| | |
|---|---|
| 1. Gutachter & Betreuer: | Dr. Stefan Schiffler |
| 2. Gutachter: | Dr. Dennis Trede |

## Abstract

In geosciences, especially in oceanography, every year more and more time series are recorded with more and more measurements. Hence, there is a need to analyse these data sets automatically or at least to preprocess them for manual analysis. One of these kind of data is the pressure measured at the bottom of the ocean. Most components in these data sets are dominated by the tides and therefore, hard to identify. In this Diploma thesis we developed a new tool for analysis of sea floor pressure data sets, Sparse Decomposition. $\ell^1$ minimisation enforces sparsity in an overcomplete dictionary which yields to physical feasible decompositions. It turned out that Sparse Decomposition outperforms other novel as well as classical decomposition tools in this application.

# Contents

# 1 Introduction

In recent years due to the development in engineering the possibilities of measuring and storing time series' have grown a lot. Nowadays large observational networks like EarthScope[1] or Neptune Canada[2] collect lots of data series 24/7, hence, there is a need to process these data automatically or at least semi-automatically.

These networks and other scientists collect lots of different data while observing the ocean. In this Diploma thesis we focus on analysis of *sea floor pressure data sets*, but in principle most of the presented notions can be applied to other kinds of time series or even to images as well. The sea floor pressure data is often measured by an ocean bottom pressure meter, see Figure 1.1. This measuring device is situated on the bottom of the ocean mostly at ridges or other interesting areas to obtain the pressure of the water column above. Of course no one is interested in this pressure itself but this pressure is an indicator of the water height above the sensor.

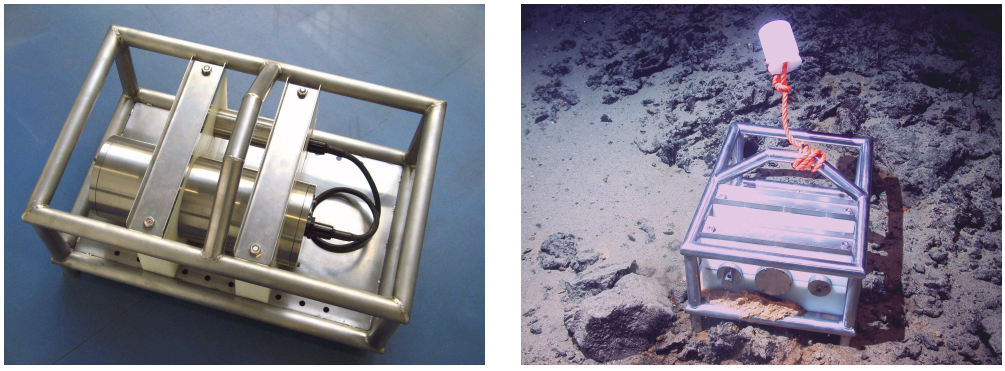


Figure 1.1.: Two different ocean bottom pressure meter for measuring the sea floor pressure. The picture on the right hand side is taken at the Logatchev Hydrothermal Field. Both are provided by Prof. Dr. Heinrich Villinger and Dr. Hans-Hermann Gennerich. © MARUM, University of Bremen

This kind of data series contains tidal pressure variations and the average pressure, but are also influenced by landslides, volcanic activity, earthquakes and

[1] `http://www.earthscope.org/`
[2] `http://www.neptunecanada.ca/`

many other sources. In general, signals small in amplitude or with a short duration are buried in large-amplitude and long-lasting signals caused by other sources. The cause of these influences is sometimes deterministic and well known as the tides, but very often especially long-term changes are not well understood. Therefore, the aim is to separate the non-deterministic signal probably small in amplitude and not visible without processing of the data from the overwhelming deterministic one.

In the past data analysis was mostly based on Fourier or Wavelet analysis but in this thesis we want to introduce the notion of *sparsity* for this application. The new approach *Sparse Decomposition* is $\ell^1$ minimisation with an overcomplete dictionary. Given a data set there are infinite many ways of representing it by an overcomplete dictionary. If our dictionary is not randomly chosen but contains pattern with physical meaning, we can seek for a representation of this data set by as less as possible pattern to achieve a physical feasible decomposition.

$\ell^1$ minimisation is in general not a new notion but the application to data analysis in geosciences is. A rather applied and good introduction to decomposition by $\ell^1$ minimisation with an overcomplete dictionary is given by Chen et al. [1999]. Major achievements in $\ell^1$ minimisation are due to Daubechies et al. [2004]. This thesis is mainly based on the recent contributions of Jin et al. [2009], Schiffler [2010] and Bredies and Lorenz [2011]. The classical variational analysis is also based on Rockafellar and Wets [1991].

To evaluate the results of Sparse Decomposition we compare them with the classical decomposition tools like Harmonic and Wavelet Decomposition, which are based on the Fourier and Wavelet transform, respectively. We also try to apply another novel decomposition tool, called Empirical Mode Decomposition, invented by Huang et al. [1998] and its enhancement the Ensemble Empirical Mode Decomposition of Wu and Huang [2009].

A crucial step in analysing data is to get familiar with the data. In Section 1.1 we present the four used data sets.

## 1.1. Sea Floor Pressure Data

Before we have a look at the single data sets, we briefly summarise some facts about water pressure. The used unit of pressure is *kilopascal* and abbreviated by *kPa*. If the circumstances, e.g. salt density and temperature, are almost constant, there is a simple relationship between the sea surface height and the sea floor pressure. Since the density of water is nearly constant in the ocean an increase of the sea floor pressure by 1 *kPa* results in an increase of almost 10 *cm* of the height of the water column above the sensor. Consequently, one might think in height of the water column [*cm*] instead of the sea floor pressure [*kPa*].

One aspect why the sea floor pressure data is interesting for geoscientists is that a change in the sea floor pressure can be the result of a vertical movement of a tectonic plate. An increase in sea floor pressure can be an increase in the water

column above the sensor and a downdrift of the tectonic plate.

Mainly, we are interested in the change of the sea floor pressure. Thus, we store our data as sea floor pressure variations around the mean of the data set.

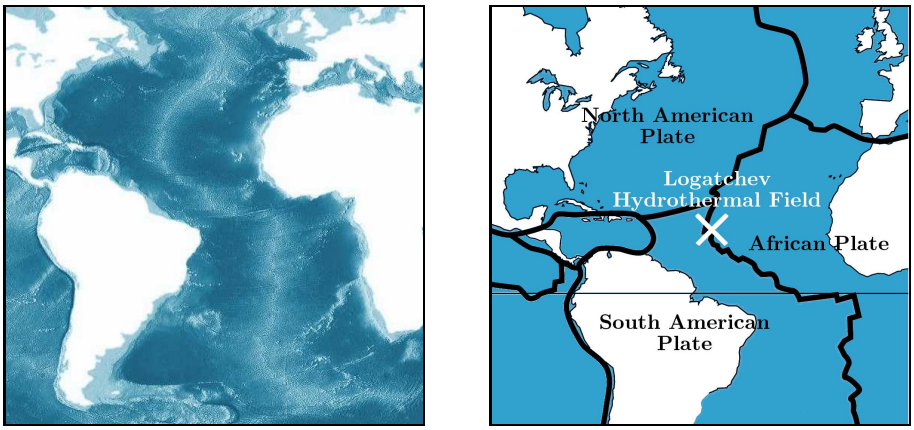



Figure 1.2.: Left: Mid-Atlantic Ridge (based on a map of NOAA[3], © NOAA); Right: Position of MAR: Logatchev Hydrothermal Field and the corresponding tectonic plates. The sensor is located at 14° 45′ N and 44° 5′ W. (based on a public domain image[4])

The first data set, called MAR, was recorded by a sensor at the Logatchev Hydrothermal Field which is located at the Mid-Atlantic Ridge, see Figure 1.2. The Mid-Atlantic Ridge is with 65,000 *km* the longest mountain range in the world even if it is mostly below the sea surface. It separates at the Logatchev Hydrothermal Field the South American Plate from the African Plate. The measurements were made to monitor the magmatic and hydrothermal activity. The duration of the data set is only one month, but due to the short sampling interval of two minutes over 22,000 measurements are needed. As a consequence of the short duration and the short sampling interval we focus on short period effects when analysing this data set.

The second and third data sets, called CORK1 and CORK2, are both recorded at Vancouver Island. The tectonic plates of interest are the small Juan de Fuca Plate and the huge surrounding plates, namely the Pacific Plate at the west end and the North American Plate at the east end. These data sets differ a lot from the first one in two features. The duration of the measurements and the number of measurements are a lot larger than at the data set MAR. CORK1 has a sampling interval of 10 minutes and by over 115,000 measurements the sea floor pressure is recorded for more than two and a half years. The data set CORK2 has a much larger sampling interval of 60 minutes but since over 9 years are recorded around

[3] `http://www.ngdc.noaa.gov/mgg/global/relief/ETOPO5/IMAGES/GIF/SLIDE15.GIF`
[4] `http://commons.wikimedia.org/wiki/File:Plaques_tectoniques_petit.gif`

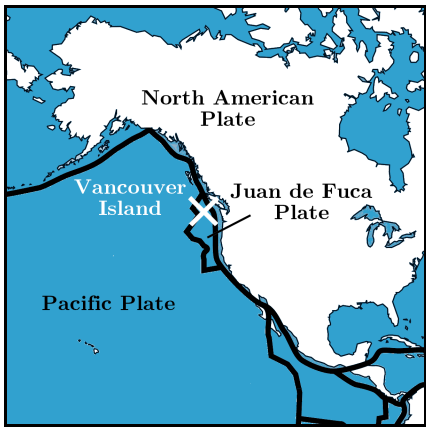



Figure 1.3: Position of CORK1 and CORK2: Vancouver Island and the corresponding tectonic plates. The sensor is located at 48° 26′ N and 128° 43′ W. (based on a public domain image[5])

80,000 measurements are needed. Another characteristic of these data sets is that the time period of CORK1 is a subset of the time period of CORK2. This might be used to evaluate boundary effects of our tools at the 'left end' of CORK1, since we know how the pressure was before starting the record.

These data sets were recorded with high-resolution absolute pressure sensors[6] with a resolution of 7 *Pa*.

The fourth data set is called SYN and is a synthetic data set. We have created this data set to evaluate the used methods. All methods have to decompose the data sets into 'real' features but since we do not know what is real there is a need for a synthetic data set with a known real decomposition.

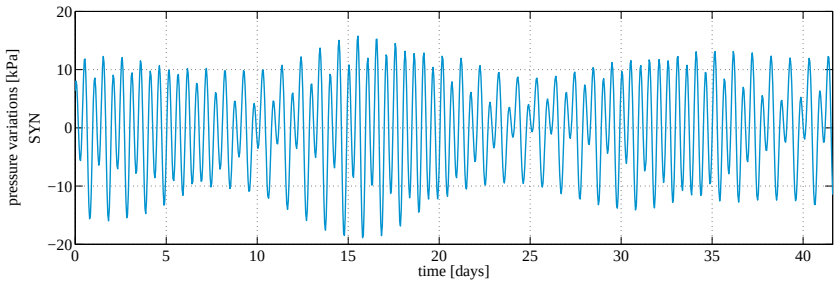



Figure 1.4.: Synthetic data set. (SYN)

The data set SYN, shown in Figure 1.4, has a length of six weeks and a sampling interval of 60 minutes, hence, 1,000 measurements are needed. It is a superposition of five different components, namely noise, a short period signal, tides, a step and monotonically increasing ramp function. These components are shown in Figure 1.5.

The two short period signals might represent earthquakes or other phenom-

[5] http://commons.wikimedia.org/wiki/File:Plaques_tectoniques_petit.gif
[6] http://www.paroscientific.com/uwapp.htm

ena with a short duration. The tides are calculated with the Matlab® function `t_tide.m` written by Pawlowicz et al. [2002]. The geographic location, which is needed for `t_tide.m`, is taken as the position of Vancouver Island. The step in the middle of the time series simulates a tectonic or oceanographic event, like a sudden downdrift of a tectonic plate, which results in an immediate decrease of the sea floor pressure. As a global trend a monotonically increasing ramp function with a total increase of roughly 0.2 $kPa$ is added. This might represent a slow but steady downdrift of the sea floor by tectonical movements. Furthermore, white noise with a RMS (root mean square) of 49.16 $Pa$ is added to the data set.

If we have a closer look at the decomposition in Figure 1.5 we see that amplitudes of the components are in a different scale. On the one hand, the tides have an amplitude of around 15 $kPa$, which is around 1.5 $m$ when we think of the water height. On the other hand, the effects we want to detect have an amplitude from 0.1 $kPa$ to 0.5 $kPa$, which are only some centimetres in change of the water height or up to 3 % of the tides. Additionally, the added noise's amplitude is also 0.1 $kPa$ which is less than 1 % when compared to the tides. However, it is between 20 % and 100 % when compared to the effects we are looking for.

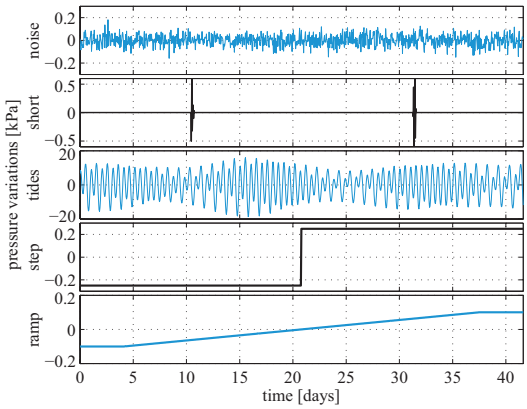



Figure 1.5.: Real decomposition of the synthetic data set.

Overall, every data set has a specialised property to evaluate the used tools. The data set MAR has a short sampling interval so this is best for detecting short lasting effects. For long time observation the data sets CORK1 and CORK2 are well suited either with a sampling interval of medium length or with an extraordinary duration. At the data set SYN we know already what we want to find since it is created by ourselves.

A summary of the details of all four data sets is given in Table 1.1.

## 1.2. Overview

First of all, we have a look at general minimisation problems and the necessary framework. Then, we show under which assumptions a minimiser uniquely exists and derive an optimal condition, which is necessary and sufficient for the minimiser. This is all done in Chapter 2.

In Chapters 3 and 4 we apply the results to a stable $\ell^1$ minimisation, called *elastic net*, and derive a fast algorithm, called *Regularised Feature Sign Search (RFSS)*, to solve this problem in finite dimensions.

Table 1.1.: Details of the used data sets.

| Data Set | Start | End | Duration [*days*] | Sampling Interval [*min*] | Number of Measurements |
|---|---|---|---|---|---|
| SYN | | | 41.7 | 60 | 1,000 |
| MAR | 01.05.2008 | 31.05.2008 | 30 | 2 | 22,319 |
| CORK1 | 26.06.2003 | 15.09.2005 | 813 | 10 | 117,012 |
| CORK2 | 11.09.1996 | 15.09.2005 | 3292 | 60 | 78,983 |

The following Chapter 5 is dedicated to the analysis of sea floor pressure data. This contains the introduction of the five used decomposition tools as well as the presentation and discussion of the results.

Lately in Chapter 6, we conclude this thesis by a brief summary and discussion of the main points.

# Foundations of Variational Calculus and Convex Analysis

This chapter should provide the reader some basics about the foundations of variational calculus and convex analysis. Moreover, we want to treat minimisation problems in general and focus on the existence of a solution and possibilities how to find them. It turns out that convexity of the corresponding function and the notion of subdifferential calculus are two key points to reach our aim but there are also other notions which are needed to get a satisfying solution.

In the whole work we treat only real vector spaces since this is needed for the application. The theory for complex vector spaces might be very similar to the real version but at least the subdifferential calculus has to be defined in another way.

## 2.1. General Definitions and Framework

In general, in minimising theory we are looking for a solution of the problem

$$\min_{x \in A} f(x)$$

with a corresponding function $f : A \to \mathbb{R}$ and $A$ any set. We call $x^* \in A$ a solution of the minimising problem if and only if $f(x^*) = \min_{x \in A} f(x) \leq f(x)\ \forall x \in A$. Later on, we refer to $x^*$ as a *minimiser*.

Sometimes it is helpful that the set where we are looking for our solution has structure like a vector space or a Hilbert space. Let $X$ be a space of the needed structure and suppose that $A \subset X$. Then we can find a numerical function $g : X \to \mathbb{R} \cup \{\infty\}$ so that $x^*$ is a solution of $\min_{x \in A} f(x)$ if and only if $x^*$ is a solution of $\min_{x \in X} g(x)$. The function $g$, which fulfils this requirement, can be defined as

$$g(x) := \begin{cases} f(x), & \text{if } x \in A \\ \infty, & \text{if } x \notin A. \end{cases}$$

As we have seen it is very useful to allow that the function we want to minimise can take the value $\infty$, thus we define some calculations as well as an ordering on $\mathbb{R} \cup \{\infty\}$.

**Definition 2.1.1 (extended real numbers).** *We extend the real numbers* $\mathbb{R}$ *to* $\mathbb{R}_\infty := \mathbb{R} \cup \{\infty\}$ *with the ordering*

$$t \leq \infty \quad \forall\, t \in \mathbb{R}_\infty \quad \textit{and} \quad t < \infty \Leftrightarrow t \in \mathbb{R}.$$

*The added summation and multiplication calculation rules are*

$$
\begin{aligned}
&1.\ \forall\, t \in \mathbb{R}_\infty : \quad t + \infty := \infty + t := \infty, \\
&2.\ \forall\, t > 0 : \qquad t \cdot \infty := \infty \cdot t \ := \infty, \\
&3. \qquad\qquad\quad 0 \cdot \infty := \infty \cdot 0 \ := 0.
\end{aligned}
$$

*Other operations like subtraction of $\infty$ and multiplication with negative real numbers are not defined.*

Note that with this ordering $\mathbb{R}_\infty$ is also a total ordered set.

In most cases we consider functions $f : X \to \mathbb{R}_\infty$. To distinguish the usual from the trivial case, i.e. $f \equiv \infty$, we define some useful terms.

**Definition 2.1.2 (effective domain, proper).** *Let $A$ be any set. For every function $f : A \to \mathbb{R}_\infty$ we define the* **effective domain** *as*

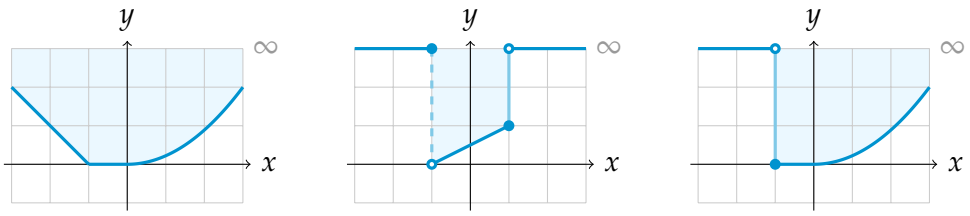
$$\operatorname{dom} f := \{x \in A : f(x) < \infty\}.$$

*$f$ is called* **proper** *if the effective domain is not empty, i.e. $\operatorname{dom} f \neq \emptyset$, which means $f \not\equiv \infty$.*

Another notion which is necessary is the *epigraph*. For the special case $f : \mathbb{R} \to \mathbb{R}$ it is the area above the graph of $f$, however, we need it for more general functions. Its topological properties give also an insight to the properties of $f$.

**Definition 2.1.3 (epigraph).** *Let $A$ be any set. For every function $f : A \to \mathbb{R}_\infty$ we define the* **epigraph** *as*

$$\operatorname{epi} f := \{(x, y) \in A \times \mathbb{R} : f(x) \leq y\}.$$




(a) closed epigraph (b) not closed epigraph (c) closed epigraph

Figure 2.1.: Three different examples of functions $f : \mathbb{R} \to \mathbb{R}_\infty$ with their epigraph coloured in light blue. The solid circle symbolises that this point belongs to the graph and the empty one does not. The light blue line indicates the boundary of the epigraph. A solid/dashed line does/does not belong to the epigraph.

To be more confident in handling functions $f : \mathbb{R} \to \mathbb{R}_\infty$ and their epigraphs three examples are shown in Figure 2.1. The function in Figure 2.1a does not take the value $\infty$, hence, the epigraph is the area over the graph. On the contrary, the other functions take the value $\infty$ and the epigraph is 'cutted' at this points.

Another mentionable fact is that the epigraph can be closed (Figure 2.1c) or not (Figure 2.1b). In general, if a function is proper the epigraph is never open.

### 2.1.1. Convexity

A pretty helpful property in variational calculus is the convexity of a function. The strict convexity provides the uniqueness of the minimiser what will be proven later on.

**Definition 2.1.4 (convex and strictly convex functions).** *Let $X$ be a vector space. The function $f : X \to \mathbb{R}_\infty$ is called* **convex** *if for all $x_1, x_2 \in X$ and $0 \leq \lambda \leq 1$*

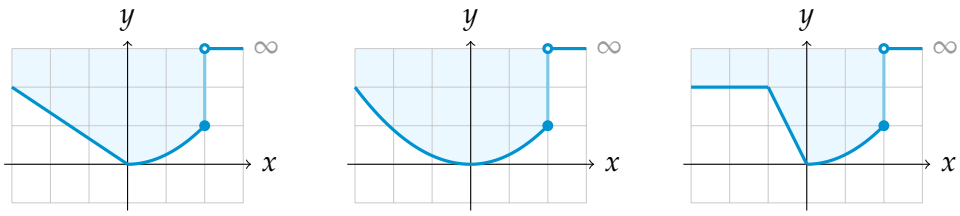
$$f(\lambda x_1 + (1-\lambda)x_2) \leq \lambda f(x_1) + (1-\lambda) f(x_2)$$

*is satisfied. It is called* **strictly convex** *if also the strict inequality*

$$f(\lambda x_1 + (1-\lambda)x_2) < \lambda f(x_1) + (1-\lambda) f(x_2)$$

*holds for every $x_1, x_2 \in X, x_1 \neq x_2, f(x_1), f(x_2) < \infty$ and $0 < \lambda < 1$.*

Some examples for convex, strictly convex and non-convex functions are given in Figure 2.2. Since the right hand side of these functions is the same the important part is the left hand side. The reason why the function in Figure 2.2a is only convex but not strictly convex is the linear part and the problem of the function in Figure 2.2c is the hook.




(a) convex (b) strictly convex (c) not convex

Figure 2.2.: The three different types of convexity: a) convex, b) strictly convex and c) not convex. The epigraphs is of these functions are drawn in light blue.

**Remark 2.1.5.** $f$ is convex if and only if epi $f$ is convex. Since dom $f$ is the image of the projection to $X$ of epi $f$ and this is a linear transformation, dom $f$ is convex as well if $f$ is convex. [Rockafellar, 1972]

This yields an alternative definition of convexity which can be visualised, at least for functions from $\mathbb{R}$ to $\mathbb{R}$.

**Lemma 2.1.6 (sum of convex functions).** *Let $X$ be a vector space and $f, g : X \to \mathbb{R}_\infty$ be two functions. Then there are the following two rules for the sum these functions.*

1. *$f, g$ convex* $\quad\Rightarrow\quad$ *$f+g$ convex*
2. *$f$ convex and $g$ strictly convex* $\quad\Rightarrow\quad$ *$f+g$ strictly convex*

*Proof.* Let $x_1, x_2 \in X$ and $0 \leq \lambda \leq 1$ be arbitrary chosen. First, we consider the case that the functions are convex. Then there is

$$\begin{aligned}(f+g)(\lambda x_1 + (1-\lambda)x_2) &= f(\lambda x_1 + (1-\lambda)x_2) + g(\lambda x_1 + (1-\lambda)x_2) \\ &\leq \lambda f(x_1) + (1-\lambda)f(x_2) + \lambda g(x_1) + (1-\lambda)g(x_2) \\ &= \lambda[f(x_1) + g(x_1)] + (1-\lambda)[f(x_2) + g(x_2)] \\ &= \lambda(f+g)(x_1) + (1-\lambda)(f+g)(x_2).\end{aligned}$$

Next, if $g$ is strictly convex, then the inequality is strict. □

**Lemma 2.1.7 (multiple of a convex function).** *Let $X$ be a vector space, $f : X \to \mathbb{R}_\infty$ and $\alpha \in \mathbb{R}$. Then the following multiplication rules hold.*

1. *$f$ convex and $\alpha \geq 0$* $\quad\Rightarrow\quad$ *$\alpha f$ convex*
2. *$f$ stricly convex and $\alpha > 0$* $\quad\Rightarrow\quad$ *$\alpha f$ strictly convex*

*Proof.* Let $x_1, x_2 \in X$ and $0 \leq \lambda \leq 1$ be arbitrary chosen.
Ad 1.: Straightforward calculations result

$$\begin{aligned}(\alpha f)(\lambda x_1 + (1-\lambda)x_2) &= \alpha f(\lambda x_1 + (1-\lambda)x_2) \\ &\leq \alpha[\lambda f(x_1) + (1-\lambda)f(x_2)] \\ &= \lambda\alpha f(x_1) + (1-\lambda)\alpha f(x_2) \\ &= \lambda(\alpha f)(x_1) + (1-\lambda)(\alpha f)(x_2).\end{aligned}$$

Ad 2.: If the multiplier is positive and the function strictly convex, then the inequation is strict. □

### 2.1.2. Dual Spaces and Weak Convergence

In many parts of this diploma thesis we consider dual spaces. Of course the reader should be familiar with dual spaces but to adjust the definition and notation these are stated in the following.

**Definition 2.1.8 (dual and double dual space).** *Let $X$ be a normed space, then $X'$ denotes the* **dual space** *of $X$ which is the space of all linear and continuous mappings from $X$ to $\mathbb{R}$, i.e. $X' := \mathcal{L}(X, \mathbb{R})$.*

*Motivated by the Riesz representation theorem 2.1.9 we denote for every $x \in X$ and $x' \in X'$ $x'(x)$ by $\langle x', x\rangle$. It looks like a scalar product, but it is <u>not</u> even if most of its properties are also valid here. Since the dual space of a normed space is again a normed space, it has a dual space. This dual space is called the* **double dual space** *of $X$, $X'' := (X')'$.*

**Theorem 2.1.9 (Riesz representation theorem, Alt [2006, page 163]).** *Let $\mathcal{H}$ be a Hilbert space. Then the mapping $J : \mathcal{H} \to \mathcal{H}', J(x)(y) := \langle x, y\rangle_{\mathcal{H}}$ is an isometric isomorphism.*

Also often used when talking about dual spaces is the adjoint operator.

**Definition 2.1.10 (adjoint operator).** *Let $X, Y$ be normed spaces and $A \in \mathcal{L}(X, Y)$. Then the* **adjoint operator** *$A'$ is a mapping $A' : Y' \to X'$ which satisfies for every $y' \in Y'$ and $x \in X$*

$$\langle A'y', x\rangle = \langle y', Ax\rangle.$$

The suitable space for our minimisation problems are reflexive spaces, since it gives us the existence of a weak limit of minimising sequences. As it turns out later on, this weak limit is a good candidate for a minimiser.

**Definition 2.1.11 (reflexive space).** *Let $X$ be a normed space. We define the mapping $\natural : X \to X''$ by $\langle \natural(x), x'\rangle := \langle x', x\rangle$. This mapping is well defined, linear, continuous and injective [Rudin, 1991, page 95]. If it is also surjective $X$ is called* **reflexive**.

**Remark 2.1.12.** Since we consider only real normed spaces, every dual space is a Banach space [Alt, 2006, page 142] and consequently every reflexive space is a Banach space as well.

The norm topology is sometimes very restrictive and to prove the existence of a minimiser it might be more appropriate to use the notion of weak convergence since there exists more weak than norm converging sequences.

**Definition 2.1.13 (weak convergence).** *Let $X$ be a normed space. We say a sequence $(x_n)_{n\in\mathbb{N}} \in X^{\mathbb{N}}$* **converges weakly** *to $x^* \in X$ if for all $x' \in X'$ there holds*

$$\lim_{n\to\infty} \langle x', x_n\rangle = \langle x', x^*\rangle$$

*and notate this by* $\text{w-lim}_{n\to\infty} x_n = x^*$.

**Remark 2.1.14.** Since there holds for every $x, x_n \in X$ and $x' \in X'$

$$|\langle x', x_n\rangle - \langle x', x\rangle| = |\langle x', x_n - x\rangle| \leq \|x'\|\|x_n - x\|_X,$$

every sequence which converges in norm converges also weakly and the weak limit is the same as the norm limit.

**Remark 2.1.15.** In Hilbert spaces $\mathcal{H}$ exist an equivalent expression of weak convergence, because their dual space is isomorph to themself, i.e. $\mathcal{H} \simeq \mathcal{H}'$, see Riesz representation theorem 2.1.9.

$$\text{w-lim}_{n\to\infty} x_n = x^* \quad \Leftrightarrow \quad \lim_{n\to\infty} \langle x, x_n\rangle_{\mathcal{H}} = \langle x, x^*\rangle_{\mathcal{H}} \quad \forall\, x \in \mathcal{H}$$

At last in this section we state two lemmas when weak convergence and convergence of some norms imply norm convergence.

**Lemma 2.1.16.** *Let $\mathcal{H}$ be any Hilbert space, $(x_n)_{n\in\mathbb{N}} \in \mathcal{H}^{\mathbb{N}}$, $x^* \in \mathcal{H}$,* w-$\lim_{n\to\infty} x_n = x^*$ *and* $\lim_{n\to\infty} \|x_n\| = \|x^*\|$. *Then there holds* $\lim_{n\to\infty} x_n = x^*$.

*Proof.* Since we are in a Hilbert space, we can use the remark as well as the usual Hilbert space identity $\|\cdot\|^2 = \langle\cdot,\cdot\rangle$, thus,

$$\begin{aligned}\|x_n - x^*\|^2 &= \|x_n\|^2 - 2\langle x^*, x_n\rangle + \|x^*\|^2 \\ &= \|x_n\|^2 - \|x^*\|^2 - 2\langle x^*, x_n\rangle + 2\langle x^*, x^*\rangle \\ &\leq \underbrace{\left|\|x_n\|^2 - \|x^*\|^2\right|}_{\to 0} + 2\underbrace{\left|\langle x^*, x^* - x_n\rangle\right|}_{\to 0}.\end{aligned}$$

□

If we do not have a Hilbert space we have to restrict ourselves to the special case $\ell^p, 1 \leq p < \infty$. By using the counting measure $\delta$ we can rewrite the norm of any $x \in \ell^p$ as

$$\|x\|_{\ell^p}^p = \sum_{n=1}^{\infty} |x_n|^p = \int_{\mathbb{N}} |x|^p \, d\delta,$$

which gives us the possibility to use Fatou's lemma.

**Theorem 2.1.17 (Fatou's lemma, Elstrodt [2004, page 144]).** *Let $(X, \mathcal{A}, \mu)$ be a measure space. Then for every sequence of non-negative measureable functions $(f^k)_{k\in\mathbb{N}}$, $f^k : X \to \mathbb{R}_\infty$, there holds*

$$\int_X \liminf_{k\to\infty} f^k \, d\mu \quad \leq \quad \liminf_{k\to\infty} \int_X f^k \, d\mu.$$

**Lemma 2.1.18 (Jin et al. [2009, page 6]).** *Let $1 \leq p < \infty$, $1 < q < \infty$, $(x^k)_{k\in\mathbb{N}} \in (\ell^q)^{\mathbb{N}}$, $x^* \in \ell^q$,* w-$\lim_{k\to\infty} x^k = x^*$ *in $\ell^q$ and* $\lim_{k\to\infty} \|x^k\|_{\ell^p} = \|x^*\|_{\ell^p}$. *Then there holds* $\lim_{k\to\infty} x^k = x^*$ *in $\ell^p$.*

*Proof.* Fatou's lemma 2.1.17 implies

$$\begin{aligned}\limsup_{k\to\infty} \|x^k - x^*\|_{\ell^p}^p &= \limsup_{k\to\infty} \left[\|x^k\|_{\ell^p}^p + \|x^*\|_{\ell^p}^p - (\|x^k\|_{\ell^p}^p + \|x^*\|_{\ell^p}^p) + \|x^k - x^*\|_{\ell^p}^p\right] \\ &\leq \limsup_{k\to\infty} \left[\|x^k\|_{\ell^p}^p + \|x^*\|_{\ell^p}^p\right] \\ &\quad + \limsup_{k\to\infty} \left[-(\|x^k\|_{\ell^p}^p + \|x^*\|_{\ell^p}^p) + \|x^k - x^*\|_{\ell^p}^p\right] \\ &= 2\|x^*\|_{\ell^p}^p - \liminf_{k\to\infty} \sum_{n=1}^{\infty} \left[|x_n^k|^p + |x_n^*|^p - |x_n^k - x_n^*|^p\right] \\ &\leq 2\|x^*\|_{\ell^p}^p - \sum_{n=1}^{\infty} \liminf_{k\to\infty} \left[|x_n^k|^p + |x_n^*|^p - |x_n^k - x_n^*|^p\right].\end{aligned}$$

Finally, since weak convergence in $\ell^q$ implies pointwise convergence, we have

$$\begin{aligned}\limsup_{k\to\infty} \|x^k - x^*\|_{\ell^p}^p &\leq 2\|x^*\|_{\ell^p}^p - \sum_{n=1}^{\infty} \liminf_{k\to\infty} \left[|x_n^k|^p + |x_n^*|^p - |x_n^k - x_n^*|^p\right] \\ &= 2\|x^*\|_{\ell^p}^p - \sum_{n=1}^{\infty} \left[|x_n^*|^p + |x_n^*|^p\right] = 2\|x^*\|_{\ell^p}^p - 2\|x^*\|_{\ell^p}^p = 0.\end{aligned}$$

□

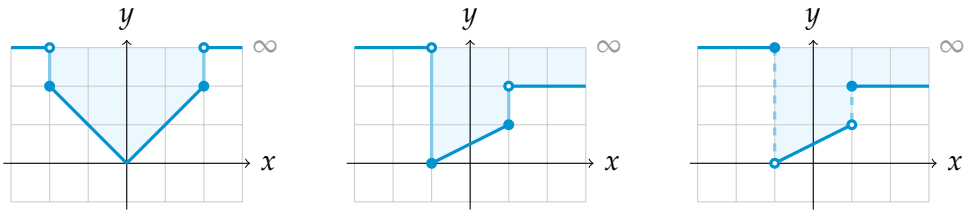
### 2.1.3. Semicontinuity

Let us have a second look at Figure 2.1 on page 8. There is an important point which was not mentioned yet. Two of the illustrated functions do have a minimiser and one has not. The function plotted in Figure 2.1a is continuous and has a minimiser but the one in Figure 2.1c is not continuous and also has a minimiser. Obviously continuity is not necessary for having a minimiser, it is lower semicontinuity or in infinite dimensions weak lower semicontinuity.

**Definition 2.1.19 (lower semicontinuity, weak lower semicontinuity).** *Let $X$ be a normed space and a function $f : X \to \mathbb{R}_\infty$. $f$ is called* **lower semicontinuous** *if for every sequence $(x_n)_{n\in\mathbb{N}} \in X^{\mathbb{N}}$ and $x \in X$ with $\lim_{n\to\infty} x_n = x$*

$$f(x) \leq \liminf_{n\to\infty} f(x_n)$$

*is fulfilled. $f$ is called* **weak lower semicontinuous** *if this is also fulfilled if the sequence $(x_n)_{n\in\mathbb{N}}$ converges only weakly.*




(a) lower semicontinuous (b) lower semicontinuous (c) not lower semicontinuous

Figure 2.3.: The solid circle symbolises that this point belongs to the graph and the open one does not. The functions in a) and b) are lower semicontinuous, but the function in c) is not. Additionally the epigraph of these functions is plotted in light blue

Here as well we want to clarify this terms by using some examples. In finite dimensions every weak convergent sequence is also norm convergent and vice versa and we are only able to plot functions $f : \mathbb{R} \to \mathbb{R}_\infty$. Thus, we can not

show the differences of these terms by plotting some examples. But what we can do is to have a look which functions are lower semicontinuous and which are not. One can easily see that in normed spaces every continuous function is lower semicontinuous, thus we have to have a look at discontinuous functions. Some examples are shown in Figure 2.3. We see that at points where the function is discontinuous and the left and right limit exists, i.e. both limits are different, the lower one has to belong to the graph.

Another fact can also be seen from Figure 2.3. The examples in Figure 2.3a and 2.3b are lower semicontinuous and have a closed epigraph. Contrary to these examples, the function in Figure 2.3c is not lower semicontinuous and its epigraph is not closed. In the following lemma we generalise this fact and prove it.

**Lemma 2.1.20 (Bredies and Lorenz [2011, page 250]).** *Let $X$ be a normed space. The function $f : X \to \mathbb{R}_\infty$ is*

1. *lower semicontinuous if and only if* epi $f$ *is closed and*
2. *weak lower semicontinuous if and only if* epi $f$ *is weak sequentially closed.*

*Proof.* We prove only the first part of the lemma. The weak version of this lemma can be proven in the same way.

Ad '$\Rightarrow$'. Let $(x_n, y_n)_{n\in\mathbb{N}} \in (\operatorname{epi} f)^{\mathbb{N}}$ with $\lim_{n\to\infty}(x_n, y_n) = (x, y)$. For every $n \in \mathbb{N}$ there is $f(x_n) \leq y_n$ since $(x_n, y_n) \in \operatorname{epi} f$. Then we have

$$f(x) \leq \liminf_{n\to\infty} f(x_n) \leq \liminf_{n\to\infty} y_n = y$$

and consequently $(x, y) \in \operatorname{epi} f$.

Ad '$\Leftarrow$'. Let $(x_n)_{n\in\mathbb{N}} \in X^{\mathbb{N}}$ be a convergent sequence with $\lim_{n\to\infty} x_n = x$. Using the notation $y_n := f(x_n)$ the sequence $(x_n, y_n)_{n\in\mathbb{N}}$ belongs to the epigraph of $f$. There is a strictly increasing mapping $\eta : \mathbb{N} \to \mathbb{N}$ so that the subsequence $(y_{\eta(n)})_{n\in\mathbb{N}}$ of $(y_n)_{n\in\mathbb{N}}$ converges to the limit inferior, i.e. $\lim_{n\to\infty} y_{\eta(n)} = \liminf_{n\to\infty} y_n$.

As the epigraph is sequentially closed, there exists again a strictly increasing mapping $\mu : \mathbb{N} \to \mathbb{N}$ and $(x^*, y^*) \in \operatorname{epi} f$ so that $\lim_{n\to\infty} x_{\mu(\eta(n))} = x^*$ and $\lim_{n\to\infty} y_{\mu(\eta(n))} = y^*$. On the one hand, the sequence $(x_n)_{n\in\mathbb{N}}$ converges to $x$ and on the other hand, it has a subsequence converging to $x^*$, hence $x = x^*$ and we obtain

$$f(x) = f(x^*) \leq y^* = \lim_{n\to\infty} y_{\mu(\eta(n))} = \lim_{n\to\infty} y_{\eta(n)} = \liminf_{n\to\infty} f(x_n).$$

□

**Remark 2.1.21.** The statements in the lemma above about lower semicontinuity and weak lower semicontinuity differ a little bit since in metric spaces a set is sequentially closed if and only if it is closed in the topology generated by the metric. However, there is no metric which generates the weak topology and this can not be used for the weak lower semicontinuity.

The last statement of this section is about the difference of lower semicontinuity and weak lower semicontinuity for convex functions. It is based on the following classical theorem.

**Theorem 2.1.22 (Werner [2007, page 108]).** *In a normed space every closed and convex set is weak sequentially closed.*

**Corollary 2.1.23.** *Let $X$ be a normed space and $f : X \to \mathbb{R}_\infty$ a convex function. Then $f$ is weak lower semicontinuous if and only if it is lower semicontiuous.*

*Proof.* Ad '$\Rightarrow$'. Since norm convergence implies weak convergence, every weak lower semicontinuous function is lower semicontiuous.

Ad '$\Leftarrow$'. Let $f$ be a lower semicontinuous function. Then epi $f$ is closed and convex and therefore weak sequentially closed, see Theorem 2.1.22. Using Lemma 2.1.20 we obtain that $f$ is weak lower semicontinuous. □

## 2.2. Existence Theorems

In the previous section of this chapter we collected lots of notions and simple properties of them which we need to state a powerful existence theorem for solutions of minimising problems in infinite dimensions. This existence theorem is called the *direct method*. Before closing this section we also have a look at finite versions of existence theorems which might have less restrictive assumptions.

### 2.2.1. The Direct Method

The existence of a minimiser in infinite dimensions is based on a classical theorem.

**Theorem 2.2.1 (Werner [2007, page 107]).** *In a reflexive space every bounded sequence has a weakly convergent subsequence.*

If we want to use this theorem then minimising sequences have to be bounded, which is clearly not satisfied in any case as we will see in the following example.

**Example 2.2.2.** Let us have a look at the minimising problem $\min_{x\in\mathbb{R}} f(x)$ with $f : \mathbb{R} \to \mathbb{R}, f(x) := \exp(x)$. The function has nearly all nice properties, it is smooth, bounded from below, strictly convex but it fails to have a minimiser because every minimising sequence is not bounded, e.g. $x_n := -n$.

The missing property of the exponential function is coercivity. This can be described for proper minimisation problems as the solution can not be found 'at infinity', but we will specify this in the following.

**Definition 2.2.3 (coercive).** *Let $X$ be a normed space and $f : X \to \mathbb{R}_\infty$ a function. $f$ is called* **coercive** *if for every sequence $(x_n)_{n\in\mathbb{N}} \in X^{\mathbb{N}}$ with $\lim_{n\to\infty} \|x_n\|_X = \infty$ there holds*

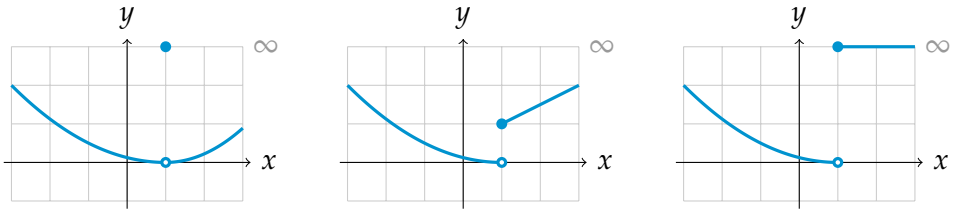
$$\lim_{n\to\infty} f(x_n) = \infty.$$

By applying this concept to functions defined on a reflexive space we get one of the most important ingredients for the direct method.

**Corollary 2.2.4 (Bredies and Lorenz [2011, page 252]).** *Let $X$ be a reflexive space, $f : X \to \mathbb{R}_\infty$ a proper and coercive function and we consider the minimising problem $\min_{x\in X} f(x)$. Then every minimising sequence has a weak convergent subsequence.*

*Proof.* Since $f$ is coercive, every minimising sequence $(x_n)_{n\in\mathbb{N}} \in X^{\mathbb{N}}$ is bounded. Assume that there exists an unbounded minimising sequence $(x_n)_{n\in\mathbb{N}} \in X^{\mathbb{N}}$, which means that $\lim_{n\to\infty} \|x_n\|_X = \infty$ and $\lim_{n\to\infty} f(x_n) = \min_{x\in X} f(x)$. But then, because of the coercivity of $f$, there is $\infty = \lim_{n\to\infty} f(x_n) = \min_{x\in X} f(x)$, which is a contradiction to the properness of $f$. Using Theorem 2.2.1 completes the proof. □

**Example 2.2.5.** With all the examples stated above it is now easy to see that also coercivity or coercivity and convexity is not sufficient for the existence of a minimiser what is shown in Figure 2.4.




(a) coercive (b) coercive, no ∞ values taken (c) coercive and convex

Figure 2.4.: In all three cases the coercive function does not have a minimiser, even if the function has any infinite values or is convex. The solid circle symbolises that this point belongs to the graph and the open one does not.

Finally, we have introduced all needed concepts for the direct method.

**Theorem 2.2.6 (direct method in reflexive spaces, Bredies and Lorenz [2011, pages 250 and 254]).** *Let $X$ be a reflexive space and $f : X \to \mathbb{R}_\infty$ a weak lower semicontinuous, coercive and proper function which is bounded from below. Then the minimisation problem $\min_{x\in X} f(x)$ has a solution in $X$.*

*Proof.* Since $f$ is bounded from below and proper, there exists a minimising sequence $(x_n)_{n\in\mathbb{N}}$ so that $\lim_{n\to\infty} f(x_n) = \inf_{x\in X} f(x) \in \mathbb{R}$.

From Corollary 2.2.4 we obtain that there exists a weak limit $x^* \in X$ and a strictly increasing mapping $\eta : \mathbb{N} \to \mathbb{N}$ so that w-$\lim_{n\to\infty} x_{\eta(n)} = x^*$. By the weak lower semicontinuity of $f$ we obtain

$$\inf_{x\in X} f(x) \leq f(x^*) \leq \liminf_{n\to\infty} f(x_{\eta(n)}) = \lim_{n\to\infty} f(x_{\eta(n)}) = \lim_{n\to\infty} f(x_n) = \inf_{x\in X} f(x),$$

thus, we proved that

$$f(x^*) = \inf_{x \in X} f(x) = \min_{x \in X} f(x).$$

This is nonetheless that $x^*$ is a minimiser of $f$. □

### 2.2.2. Existence Theorems in Finite Dimensions

In real world applications, especially when using the computer, no one has infinite dimensional spaces. Therefore, we might not need such restrictive conditions for the existence of a minimiser. The requirement of properness is obviously necessary in any case and can not be skipped. Also the coercivity is necessary, as we have seen in Example 2.2.2. But let us see if the other conditions might be relaxed and have a look at the work of Rockafellar and Wets [1991].

**Theorem 2.2.7 (minimisation in finite dimensional spaces, Rockafellar and Wets [1991, page 11]).** *Let $f : \mathbb{R}^M \to \mathbb{R}_\infty$ be a lower semicontinuous, coercive and proper function. Then the minimisation problem $\min_{x \in X} f(x)$ has a solution in $\mathbb{R}^M$.*

In finite dimensional spaces weak lower semicontinuity is the same as lower semicontinuity as a sequence is weak convergent if and only if it is norm convergent. Only the boundedness from below might be skipped since it comes automically from the other conditions using the Heine-Borel theorem.

To complete this section we add a third version of an existence theorem for minimisation problems. In this case we consider convex minimisation problems and restrict our function to have only finite values. This constraint does not look very restrictive but as we see in the following it is quite powerful.

**Theorem 2.2.8 (convex minimisation in finite dimensional spaces).** *Let $f : \mathbb{R}^M \to \mathbb{R}$ be a convex and coercive function. Then the minimisation problem $\min_{x \in X} f(x)$ has a solution in $\mathbb{R}^M$.*

*Proof.* We want to apply Theorem 2.2.7. First of all, our function $f$ is obviously proper since it is not allowed to have the value $\infty$ at any point. Secondly finite, convex functions are continuous, see Rockafellar and Wets [1991, page 61]. Therefore, $f$ fulfils all conditions of Theorem 2.2.7. □

By comparing this theorem to a similar one provided by Bredies and Lorenz [2011, page 263] for infinite dimensional convex minimisation problems, we see that we get this time the lower semicontinuity for free.

## 2.3. Differential Calculus

### 2.3.1. Differential Calculus in Infinite Dimensions

Differentiability is an important property in the theory of minimisation problems. At least for smooth functions $f : \mathbb{R} \to \mathbb{R}$ it is known, that a vanishing derivative

is a necessary condition for the minimiser. In the following we want to state two notions for differentiability in infinite dimensional spaces which are called *Gâteaux differentiability* and *Fréchet differentiability*. To complete the introduction we state some properties of these derivatives and give some simple examples which are used in the next chapter. This section is mainly based on Schiffler [2010, pages 16 and 17] and Werner [2007, pages 112-126].

**Definition 2.3.1 (Gâteaux differentiability).** *Let $X, Y$ be normed spaces and $A \subset X$ an open subset. The mapping $f : A \to Y$ is called* **Gâteaux differentiable at a point** *$x^* \in U$ if there exists a mapping $K \in \mathcal{L}(X, Y)$ so that for every $x \in X$*

$$\lim_{h \to 0} \| \tfrac{1}{h} [f(x^* + hx) - f(x^*)] - Kx \|_Y = 0. \tag{2.3.1}$$

*We denote $K$ by $\mathcal{G}_f(x^*)$ and call it the* **Gâteaux derivative** *of $f$ in $x^*$. $f$ is called* **Gâteaux differentiable** *if it is Gâteaux differentiable at any point $x^* \in A$.*

**Proposition 2.3.2 (properties of Gâteaux differentiability, Werner [2007, page 120,121]).** *Let $X, Y$ be normed spaces, $A \subset X$ an open subset and $f, g : A \to Y$ are Gâteaux differentiable at $x^* \in A$.*

1. ***sum rule:*** *The mapping $f + g$ is Gâteaux differentiable at $x^* \in A$ and the derivative is given by*

$$\mathcal{G}_{f+g}(x^*) = \mathcal{G}_f(x^*) + \mathcal{G}_g(x^*).$$

2. ***scalar multiplication:*** *For any $\lambda \in \mathbb{R}$ the mapping $\lambda f$ is Gâteaux differentiable at $x^* \in A$ and the derivative is given by*

$$\mathcal{G}_{\lambda f}(x^*) = \lambda \mathcal{G}_f(x^*).$$

**Theorem 2.3.3 (minimisation with Gâteaux differentiability, Werner [2007, page 123]).** *Let $X$ be a normed space, $A \subset X$ an open subset and $f : A \to \mathbb{R}$ Gâteaux differentiable. If $f$ has a local minima at $x^* \in U$, i.e. there exists an open set $B \subset A$ so that $x^* \in B$ and for every $x \in B$ there holds $f(x^*) \leq f(x)$, then there is $\mathcal{G}_f(x^*) = 0$.*

We have seen that Gâteaux differentiability is a nice notion to have a derivative in infinite dimensions which is similar to the finite dimensional one. But one of the most important equations is not valid, the *chain rule*. The chain rule holds for *Fréchet differentiable* mappings.

**Definition 2.3.4 (Fréchet differentiability).** *Let $X, Y$ be normed spaces and $A \subset X$ an open subset. The mapping $f : A \to Y$ is called* **Fréchet differentiable at a point** *$x^* \in A$ if there exists a mapping $K \in \mathcal{L}(X, Y)$ so that the convergence as in equation* (2.3.1) *is uniform for $\|x\|_X \leq 1$, i.e.*

$$\lim_{h \to 0} \| \tfrac{1}{h} [f(x^* + h\cdot) - f(x^*)] - K\cdot \|_\infty = 0.$$

*We denote K by $\mathcal{F}_f(x^*)$ and call it the* **Fréchet derivative** *of $f$ in $x^*$. $f$ is called* **Fréchet differentiable** *if it is Fréchet differentiable at any point $x^* \in A$.*

The norm, which is used in the definition above, is the supremum norm on the closed unit ball $B_1(0) \subset X$. It is defined as $\|g\|_\infty := \sup_{x \in B_1(0)} \|g(x)\|_Y$ for any function $g : B_1(0) \to Y$ .

**Remark 2.3.5.** Obviously, every Fréchet differentiable mapping is also Gâteaux differentiable. On the contrary, a Gâteaux differentiable mapping might fail to be Fréchet differentiable.

**Proposition 2.3.6 (chain rule for Fréchet differentiability, Werner [2007, pages 120 and 121]).** *Let $X, Y, Z$ be normed spaces and $A \subset X, B \subset Y$ open subsets. If $f : A \to Y$ with $f(A) \subset B$ is Fréchet differentiable at $x^* \in A$ and $g : B \to Z$ is Fréchet differentiable at $f(x^*) \in B$ then the mapping $g \circ f : A \to Z$ is Fréchet differentiable at $x^*$ and the Fréchet derivative is given by*

$$\mathcal{F}_{g \circ f}(x^*) = \mathcal{F}_g(f(x^*)) \circ \mathcal{F}_f(x^*).$$

**Example 2.3.7.** Let us consider two normed spaces $X, Y$ and an operator $K \in \mathcal{L}(X, Y)$. Then the Fréchet derivative of $K$ at any point $x^* \in X$ is $\mathcal{F}_K(x^*) = K$, because

$$\begin{aligned} \|\tfrac{1}{h}[K(x^* + hx) - Kx^*] - \mathcal{F}_K(x^*)x\|_Y &= \|\tfrac{1}{h}[Kx^* + hKx - Kx^*] - Kx\|_Y \\ &= \|0\|_Y = 0. \end{aligned}$$

**Example 2.3.8.** As a second example we calculate the Fréchet derivative of $\|\cdot\|_{\mathcal{H}}^2 : \mathcal{H} \to \mathbb{R}$, where $\mathcal{H}$ is a Hilbert space. Using the Hilbert space identity we obtain

$$\begin{aligned} \tfrac{1}{h}\left[\|x^* + hx\|_{\mathcal{H}}^2 - \|x^*\|_{\mathcal{H}}^2\right] &= \tfrac{1}{h}\left[\|x^*\|_{\mathcal{H}}^2 + 2h\langle x^*, x\rangle_{\mathcal{H}} + h^2\|x\|_{\mathcal{H}}^2 - \|x^*\|_{\mathcal{H}}^2\right] \\ &= 2\langle x^*, x\rangle_{\mathcal{H}} + h\|x\|_{\mathcal{H}}^2 \xrightarrow{h \to 0} 2\langle x^*, x\rangle_{\mathcal{H}} \end{aligned}$$

and so we get the Gâteaux derivative $\mathcal{G}_{\|\cdot\|_{\mathcal{H}}^2}(x^*) = 2\langle x^*, \cdot\rangle_{\mathcal{H}}$. But we also want to know if this is the Fréchet derivative or not.

$$\sup_{\|x\|_{\mathcal{H}} \leq 1} |\tfrac{1}{h}\left[\|x^* + hx\|_{\mathcal{H}}^2 - \|x^*\|_{\mathcal{H}}^2\right] - 2\langle x^*, x\rangle_{\mathcal{H}}| = \sup_{\|x\|_{\mathcal{H}} \leq 1} |h\|x\|_{\mathcal{H}}^2| \leq |h| \xrightarrow{h \to 0} 0$$

Hence, the Fréchet derivative of $\|\cdot\|_{\mathcal{H}}^2$ at any point $x^*$ is given by $\mathcal{F}_{\|\cdot\|_{\mathcal{H}}^2}(x^*) = 2\langle x^*, \cdot\rangle_{\mathcal{H}}$.

**Example 2.3.9.** The third example is slightly more advanced. For a normed space $X$, a Hilbert space $\mathcal{H}$, an operator $K \in \mathcal{L}(X, \mathcal{H})$ and $y \in \mathcal{H}$ we calculate the Fréchet derivative of $\frac{1}{2}\|K(\cdot) - y\|_{\mathcal{H}}^2 : X \to \mathbb{R}$. Using the rules from Proposition 2.3.2, the chain rule from Proposition 2.3.6, both of the examples above and the fact that the

Fréchet derivative of a constant function is zero, we immediately get the Fréchet derivative at any $x^* \in X$ as

$$\begin{aligned}\mathcal{F}_{\frac{1}{2}\|K\cdot - y\|_{\mathcal{H}}^2}(x^*) &= \tfrac{1}{2}\mathcal{F}_{\|\cdot\|_{\mathcal{H}}^2 \circ [K(\cdot)-y]}(x^*) \\ &= \tfrac{1}{2}[\mathcal{F}_{\|\cdot\|_{\mathcal{H}}^2}(Kx^* - y) \circ \mathcal{F}_{K(\cdot)-y}(x^*)] \\ &= \tfrac{1}{2}[2\langle Kx^* - y, \cdot\rangle_{\mathcal{H}} \circ K(\cdot)] = \langle Kx^* - y, K(\cdot)\rangle_{\mathcal{H}} \\ &= \langle K'(Kx^* - y), \cdot\rangle_{\mathcal{H}}.\end{aligned}$$

### 2.3.2. Subdifferential Calculus

The aim of this section is to provide an optimality condition for solving minimisation problems. If we consider a function $f : X \to \mathbb{R}$ which is Gâteaux differentiable, a necessary condition for solving the minimisation problem $\min_{x\in X} f(x)$ is given by $\mathcal{G}_f(x) = 0$, see Theorem 2.3.3. But in our application the function is not smooth, so this notion is not applicable. Therefore, we need another notion of differentiability which leads us to a slightly different optimality condition.

Let us consider a differentiable function $f : \mathbb{R}^N \to \mathbb{R}$. $f$ is convex if and only if for all $x, x^* \in \mathbb{R}^N$ there holds

$$f(x) \geq f(x^*) + \langle \nabla f(x^*), x - x^*\rangle, \tag{2.3.2}$$

see Rockafellar and Wets [1991, page 47]. Since we consider convex functions which might fail to be differentiable, equation (2.3.2) might be a proper condition for a new notion of differentiability, called the *subdifferential calculus*. In this notion the differential might not be unique, thus, we need the notion of *set valued functions*.

**Definition 2.3.10 (set valued functions).** *Let $A$ and $B$ be any sets. The function $f : A \to \mathfrak{P}(B)$ is called a* **set valued function**, *where $\mathfrak{P}(B)$ denotes the powerset of $B$. In the following we write only $f : A \rightrightarrows B$.*

Since the values of these functions are sets, we need to have some operations on sets.

**Definition 2.3.11 (addition and scalar multiplication of sets).** *Let $X$ be a vector space. For two sets $A, B \subset X$ and $\lambda \in \mathbb{R}$ we define*

$$A + B := \{x + y \in X : x \in A, y \in B\} \quad \textit{and} \quad \lambda \cdot A := \{\lambda \cdot x \in X : x \in A\}.$$

With this definition, we can also define the addition and scalar multiplication of set valued functions. As for usual functions these are defined 'pointwise'.

**Definition 2.3.12 (addition and scalar multiplication of set valued functions).** *Let $A$ be a set and $Y$ a vector space. For set valued functions $f, g : A \rightrightarrows Y$ and $\lambda \in \mathbb{R}$ we*

*define*

$$f+g : A \rightrightarrows Y \qquad \text{and} \qquad \lambda f : A \rightrightarrows Y$$
$$x \mapsto f(x)+g(x) \qquad\qquad x \mapsto \lambda f(x).$$

**Definition 2.3.13 (subgradient, subdifferential).** *Let $X$ be a normed space and $f : X \to \mathbb{R}_\infty$ a convex function. $w^* \in X'$ is called the* **subgradient** *of $f$ at a point $x^* \in X$ if for every $x \in X$ there holds*

$$f(x) \geq f(x^*) + \langle w^*, x - x^* \rangle. \tag{2.3.3}$$

*The* **subdifferential** *$\partial f$ of $f$ is the set valued function $\partial f : X \rightrightarrows X'$ with*

$$\partial f(x^*) = \left\{ w^* \in X' : w^* \text{ is a subgradient of } f \text{ in } x^* \right\}.$$

**Remark 2.3.14.** Since equation (2.3.3) is trivial for $x \notin \operatorname{dom} f$, it is sufficient to claim this for all $x \in \operatorname{dom} f$.

**Remark 2.3.15.** The subdifferential might be multi-valued, single-valued or empty. If it is single-valued we do not differ between $\partial f(x^*) = \{w^*\}$ and $w^*$. When we have a closer look at equation (2.3.3) for $x^* \notin \operatorname{dom} f$, we see that this can only be fulfilled if $f$ is not proper. Hence, for a proper $f$ we have for $x^* \notin \operatorname{dom} f$ that $\partial f(x^*) = \emptyset$.

**Remark 2.3.16.** If we consider a Hilbert space $\mathcal{H}$, we can identify by Riesz representation theorem 2.1.9 the subgradient $\partial f(x^*)$ with

$$\{ w^* \in \mathcal{H} : f(x) \geq f(x^*) + \langle w^*, x - x^* \rangle_{\mathcal{H}} \, \forall \, x \in \mathcal{H} \},$$

where now $\langle \cdot, \cdot \rangle_{\mathcal{H}}$ denotes the scalar product of $\mathcal{H}$. Therefore, we do not differ between these sets if $\mathcal{H}$ is a Hilbert space.

For later purposes we need to define the *multi-valued sign function*. To emphasise the difference between the ordinary sign function and its multi-valued version we state the common definition of the sign function beforehand, which reads $\operatorname{sign} : \mathbb{R} \to \mathbb{R}$

$$\operatorname{sign}(t) := \begin{cases} 1, & \text{for } t > 0 \\ 0, & \text{for } t = 0 \\ -1, & \text{for } t < 0. \end{cases}$$

**Definition 2.3.17 (multi-valued sign function).** *We define the* **multi-valued sign function** *for real numbers $\mathcal{S}^1 : \mathbb{R} \rightrightarrows \mathbb{R}$ as*

$$\mathcal{S}^1(t) := \begin{cases} \{\operatorname{sign}(t)\}, & \text{for } t \neq 0 \\ [-1, 1], & \text{for } t = 0 \end{cases}$$

*and the multi-valued sign function for sequences $\mathcal{S}^{\mathbb{N}} : \mathbb{R}^{\mathbb{N}} \to \mathfrak{P}(\mathbb{R})^{\mathbb{N}}$ as*

$$\mathcal{S}^{\mathbb{N}}((x_n)_{n \in \mathbb{N}}) := (\mathcal{S}^1(x_n))_{n \in \mathbb{N}}.$$

To see why we need this and to get familiar with the notion of subdifferentials we calculate an example.

**Example 2.3.18.** Lets calculate the subdifferential of $|\cdot| : \mathbb{R} \to \mathbb{R}$ at any $x^* \in \mathbb{R}$. The condition which needs to be fulfilled reads for $x^* = 0, x \neq 0$

$$|x| \geq w^* \cdot x \Leftrightarrow 1 \geq w^* \cdot \operatorname{sign}(x).$$

Therefore, we obtain $\partial|0| = \{w^* \in \mathbb{R} : -1 \leq w^* \leq 1\} = [-1, 1]$.

For $x^* \neq 0$ we define $w^* := \operatorname{sign}(x^*) + \varepsilon$ and show that $\varepsilon = 0$. For all $x \in \mathbb{R}$ the condition is

$$|x| \geq |x^*| + (\operatorname{sign}(x^*) + \varepsilon)(x - x^*) = \operatorname{sign}(x^*)x + \varepsilon(x - x^*). \tag{2.3.4}$$

This is obviously fulfilled for $\varepsilon = 0$. Lets consider now $\varepsilon \neq 0$. Assume that the condition is fulfilled and choose

$$\tilde{x} := x^* + \tfrac{1}{2}\operatorname{sign}(\varepsilon)|x^*| = [\operatorname{sign}(x^*) + \tfrac{1}{2}\operatorname{sign}(\varepsilon)]|x^*|.$$

Then, $\operatorname{sign}(\tilde{x}) = \operatorname{sign}(x^*)$ and for $\tilde{x}$ the condition (2.3.4) reads

$$|\tilde{x}| \geq \operatorname{sign}(\tilde{x})\tilde{x} + \tfrac{1}{2}\varepsilon \operatorname{sign}(\varepsilon)|x^*| = |\tilde{x}| + \tfrac{1}{2}|\varepsilon||x^*| > |\tilde{x}|$$

which is a contradiction. Therefore, the subgradient of the absolute value at any $x^* \neq 0$ is $\partial|x^*| = \{\operatorname{sign}(x^*)\}$. Summarising the results we get for every $x^* \in \mathbb{R}$

$$\partial|x^*| = \mathcal{S}^1(x^*). \tag{2.3.5}$$

The next theorem is very import, even if the proof is rather trivial. It gives us the possibility to describe the solutions of a convex minimisation problem.

**Theorem 2.3.19 (optimality condition for convex minimisation problems, Bredies and Lorenz [2011, page 272]).** *Let $X$ be a normed space and $f : X \to \mathbb{R}_\infty$ a convex function. Then there holds for every $x^* \in X$*

$$0 \in \partial f(x^*) \Leftrightarrow x^* \in X \text{ solves the minimisation problem } \min_{x \in X} f(x).$$

*Proof.*

$$\begin{aligned} x^* \in X \text{ solves } \min_{x \in X} f(x). &\Leftrightarrow f(x^*) = \min_{x \in X} f(x) \Leftrightarrow \forall\, x \in X : f(x^*) \leq f(x) \\ &\Leftrightarrow \forall\, x \in X : f(x^*) + \langle 0, x - x^* \rangle \leq f(x) \Leftrightarrow 0 \in \partial f(x^*) \end{aligned}$$

□

Next, we want to state and prove some calculation rules that helps us to calculate the subdifferentials of more advanced and complicated functions.

First of all, we need a classical theorem because it is necessary for the next proof. This theorem is a simple consequence of the Hahn-Banach theorem

**Theorem 2.3.20 (Hahn-Banach separation theorem, Werner [2007, page 103]).** *Let $X$ be a normed space, $A_1, A_2 \subset X$ convex sets, $A_1$ open and $A_1 \cap A_2 = \emptyset$. Then there exists $x' \in X'$ so that for every $x_1 \in A_1$ and $x_2 \in A_2$*

$$x'(x_1) < x'(x_2)$$

*is fulfilled.*

**Remark 2.3.21.** This theorem can be extended to $\overline{A}_1$, but then the inequality is not strict anymore, i.e. for every $x_1 \in \overline{A}_1$ and $x_2 \in A_2$ there holds $x'(x_1) \leq x'(x_2)$.

**Remark 2.3.22.** If $A_1$ and $A_2$ are non-empty sets, then there exists $\lambda \in \mathbb{R}$ so that $x'(x_1) \leq \lambda \leq x'(x_2)$ for every $x_1 \in A_1$ and $x_2 \in A_2$, for instance $\lambda := \sup_{x_1 \in A_1} x'(x_1)$.

**Proposition 2.3.23 (calculus rules for subdifferentials, Bredies and Lorenz [2011, page 279]).** *Let $X, Y$ denote normed spaces, $f, f_1, f_2 : X \to \mathbb{R}_\infty$ proper, convex functions and $\lambda > 0$. Then there are the following rules for calculating subdifferentials.*

1. $\partial(\lambda f) = \lambda \partial f$
2. $\partial(f_1 + f_2) \supset \partial f_1 + \partial f_2$
3. *If $f_1$ is continuous in any $x_0 \in \operatorname{dom} f_1 \cap \operatorname{dom} f_2$ :* $\quad \partial(f_1 + f_2) \subset \partial f_1 + \partial f_2$

*Proof.* Ad 1.: For every $\lambda > 0$ and $x^* \in X$ there is

$$\begin{aligned} w \in \partial(\lambda f)(x^*) &\Leftrightarrow \lambda f(x) \geq \lambda f(x^*) + \langle w, x - x^* \rangle \quad \forall\, x \in \operatorname{dom} f \\ &\Leftrightarrow f(x) \geq f(x^*) + \langle \lambda^{-1} w, x - x^* \rangle \quad \forall\, x \in \operatorname{dom} f \\ &\Leftrightarrow \lambda^{-1} w \in \partial f(x^*) \Leftrightarrow w \in \lambda \partial f(x^*). \end{aligned}$$

Ad 2.: For every $x^* \in X$, $w_1 \in \partial f_1(x^*)$, $w_2 \in \partial f_2(x^*)$ and $x \in \operatorname{dom} f_1 \cap \operatorname{dom} f_2$ there holds

$$\begin{aligned} (f_1 + f_2)(x) &= f_1(x) + f_2(x) \geq f_1(x^*) + \langle w_1, x - x^* \rangle + f_2(x^*) + \langle w_2, x - x^* \rangle \\ &= (f_1 + f_2)(x^*) + \langle w_1 + w_2, x - x^* \rangle \end{aligned}$$

which is the condition for $w_1 + w_2 \in \partial(f_1 + f_2)(x^*)$.

Ad 3.: Let $x^* \in X$ be an arbitrary chosen point and $w^* \in \partial(f_1 + f_2)(x^*)$. Hence,

$$f_1(x) + f_2(x) \geq f_1(x^*) + f_2(x^*) + \langle w^*, x - x^* \rangle \quad \forall\, x \in \operatorname{dom} f_1 \cap \operatorname{dom} f_2.$$

So $x^* \in \operatorname{dom} f_1 \cap \operatorname{dom} f_2$ and we can rewrite this condition for every $x \in \operatorname{dom} f_1 \cap \operatorname{dom} f_2$ as

$$f_1(x) - f_1(x^*) - \langle w^*, x - x^* \rangle \geq f_2(x^*) - f_2(x). \tag{2.3.6}$$

By using the auxiliary function $\tilde{f}_1(x) := f_1(x) - \langle w^*, x\rangle$, which is still convex and continuous in $x_0$, we can rewrite equation (2.3.6) and get for every $x \in \operatorname{dom} f_1 \cap \operatorname{dom} f_2$

$$\tilde{f}_1(x) - \tilde{f}_1(x^*) \geq f_2(x^*) - f_2(x). \tag{2.3.7}$$

We want to find $w_2 \in X'$ which 'fits between' these terms. Let us consider the sets $A_1, A_2 \subset X \times \mathbb{R}$ with

$$A_1 := \{(x,t) : \tilde{f}_1(x) - \tilde{f}_1(x^*) \leq t\} = \{(x, t - \tilde{f}_1(x^*)) : \tilde{f}_1(x) \leq t\}$$

and

$$A_2 := \{(x,t) : t \leq f_2(x^*) - f_2(x)\} = \{(x, f_2(x^*) - t) : f_2(x) \leq t\}.$$

We would like to use the Hahn-Banach separation theorem 2.3.20 with the Remark 2.3.21 for $\mathring{A}_1$ and $A_2$. We have to verify if they are convex and disjoint. Additionally, we check the non-emptiness, because we want to use Remark 2.3.22 as well.

First, we prove the convexity of $A_1$. For every $(x_1,t_1), (x_2,t_2) \in A_1$ and $\lambda \in [0,1]$ there is

$$\begin{aligned}\tilde{f}_1(\lambda x_1 + (1-\lambda)x_2) - \tilde{f}_1(x^*) &\leq \lambda\tilde{f}_1(x_1) + (1-\lambda)\tilde{f}_1(x_2) - \tilde{f}_1(x^*)\\ &= \lambda[\tilde{f}_1(x_1) - \tilde{f}_1(x^*)] + (1-\lambda)[\tilde{f}_1(x_2) - \tilde{f}_1(x^*)]\\ &\leq \lambda t_1 + (1-\lambda)t_2,\end{aligned}$$

hence, $\lambda(x_1,t_1) + (1-\lambda)(x_2,t_2) \in A_1$, which is nonetheless as the convexity of $A_1$. The proof of the convexity of $A_2$ is analogous. Also very easy to verify is the convexity of the interior of a convex set.

Next, we check the disjointness of these sets. Assume there exists $(x,t) \in \mathring{A}_1 \cap A_2$. But then there holds

$$\tilde{f}_1(x) - \tilde{f}_1(x^*) < t \leq f_2(x^*) - f_2(x),$$

which is a contradiction to equation (2.3.7).

The third condition to check is the non-emptiness. For every $x \in X$ there is $(x, \tilde{f}_1(x) - \tilde{f}_1(x^*)) \in A_1$ and $(x, f_2(x^*) - f_2(x)) \in A_2$.

Finally, we have to prove that $\mathring{A}_1$ is not empty. For any $\varepsilon > 0$ there is $(x_0, z_0) \in A_1$, where we have used $z_0 := \tilde{f}_1(x_0) - \tilde{f}_1(x^*) + \varepsilon$. Since $\tilde{f}$ is continuous in $x_0$, there exists $\tilde{\delta} > 0$ so that $|\tilde{f}_1(x) - \tilde{f}_1(x_0)| < \varepsilon/2$ as long as $\|x - x_0\|_X < \tilde{\delta}$. With $\delta := \min(\tilde{\delta}, \varepsilon/2)$ we have for every $(x, t - \tilde{f}_1(x^*)) \in B_\delta(x_0, z_0)$ that

$$\begin{aligned}\|x_0 - x\| + |\tilde{f}_1(x_0) + \varepsilon - t| &= \|x - x_0\| + |[t - \tilde{f}_1(x^*)] - [\tilde{f}_1(x_0) - \tilde{f}_1(x^*) + \varepsilon]|\\ &= \|x - x_0\| + |[t - \tilde{f}_1(x^*)] - z_0| < \delta \leq \varepsilon/2\end{aligned}$$

and in particular

$$|\tilde{f}_1(x_0) + \varepsilon - t| \leq \|x_0 - x\| + |\tilde{f}_1(x_0) + \varepsilon - t| < \varepsilon/2.$$

Consequently, we have $(x, t - \tilde{f}_1(x^*)) \in A_1$ since

$$\begin{aligned}\tilde{f}_1(x) &\leq |\tilde{f}_1(x) - \tilde{f}_1(x_0)| + \tilde{f}_1(x_0) \\ &\leq \underbrace{|\tilde{f}_1(x) - \tilde{f}_1(x_0)|}_{<\varepsilon/2} + \underbrace{|\tilde{f}_1(x_0) + \varepsilon - t|}_{<\varepsilon/2} - \varepsilon + t < \varepsilon/2 + \varepsilon/2 - \varepsilon + t = t\end{aligned}$$

and hence we have proven that $\mathring{A}_1 \neq \emptyset$ as $(x_0, z_0) \in \mathring{A}_1$.

By applying Hahn-Bach separation theorem 2.3.20 there exist $(x_0', t_0') \in X' \times \mathbb{R}$ and $\lambda \in \mathbb{R}$ so that for every $(x_1, t_1) \in A_1$ and $(x_2, t_2) \in A_2$ there is

$$\langle x_0', x_1\rangle + t_0'[t_1 - \tilde{f}_1(x_1)] \leq \lambda \leq \langle x_0', x_2\rangle + t_0'[f_2(x_2) - t_2].$$

This can be written as

$$\langle x_0', x\rangle + t_0'[t - \tilde{f}_1(x^*)] \leq \lambda \quad \forall\, x \in \operatorname{dom} \tilde{f}_1, \tilde{f}_1(x) \leq t \tag{2.3.8}$$

$$\text{and} \qquad \langle x_0', x\rangle + t_0'[f_2(x^*) - t] \geq \lambda \quad \forall\, x \in \operatorname{dom} f_2, f_2(x) \leq t. \tag{2.3.9}$$

Now, we want to show that $t_0' < 0$. This is done by contradicting the other cases to the inequalities (2.3.8) and (2.3.9). Assume that $t_0' > 0$. But then for $x = x_0$ and large $t$, i.e. $t > \tilde{f}_1(x^*)$, we have

$$\lim_{t\to\infty} \langle x_0', x_0\rangle + t_0'[t - \tilde{f}_1(x^*)] = \infty \not\leq \lambda. \qquad \lightning$$

Assume now that $t_0' = 0$. But since $(x_0, z_0) \in \mathring{A}_1$ and $(x_0, f_2(x^*) - f_2(x_0)) \in A_2$ we have the strict inequality by Hahn-Banach separation theorem 2.3.20

$$\langle x_0', x_0\rangle = \langle (x_0', t_0'), (x_0, z_0)\rangle < \langle (x_0', t_0'), (x_0, f_2(x^*) - f_2(x_0))\rangle = \langle x_0', x_0\rangle. \qquad \lightning$$

This was the major work. After plugging in $t = \tilde{f}_1(x)$ in inequality (2.3.8) as well as $t = f_2(x)$ in inequality (2.3.9) we get

$$\tilde{f}_1(x) - \tilde{f}_1(x^*) \geq \tfrac{1}{t_0'}[-\langle x_0', x\rangle + \lambda] \quad \forall\, x \in \operatorname{dom} f_1 \tag{2.3.10}$$

$$\text{and} \qquad f_2(x^*) - f_2(x) \leq \tfrac{1}{t_0'}[-\langle x_0', x\rangle + \lambda] \quad \forall\, x \in \operatorname{dom} f_2. \tag{2.3.11}$$

But this is also true for $x^*$, since $x^* \in \operatorname{dom} f_1 \cap \operatorname{dom} f_2$. Hence, we get $\lambda = \langle x_0', x^*\rangle$. Using $w_2 := \frac{1}{t_0'} x_0'$ we obtain from inequality (2.3.11)

$$f_2(x^*) - f_2(x) \leq -\langle w_2, x - x^*\rangle \quad \forall\, x \in \operatorname{dom} f_2,$$

which is equivalent to $w_2 \in \partial f_2(x^*)$. On the other hand with $w_1 := w^* - w_2$ we get from inequality (2.3.10)

$$\begin{aligned} f_1(x) - f_1(x^*) &= \tilde{f}_1(x) - \tilde{f}_1(x^*) + \langle w^*, x - x^* \rangle \\ &\geq -\langle w_2, x - x^* \rangle + \langle w^*, x - x^* \rangle = \langle w_1, x - x^* \rangle \quad \forall\, x \in \operatorname{dom} f_1, \end{aligned}$$

which is means $w_1 \in \partial f_1(x^*)$. Concluding we have $w^* = w_1 + w_2 \in \partial f_1(x^*) + \partial f_2(x^*)$. □

**Proposition 2.3.24 (subgradient versus gradient, Bredies and Lorenz [2011, page 266]).** *Let $X$ be a normed space and $f : X \to \mathbb{R}_\infty$ a convex function. If $f$ is Gâteaux differentiable in $x^* \in \overset{\circ}{\operatorname{dom}} f$ then there is*

$$\partial f(x^*) = \{\mathcal{G}_f(x^*)\}.$$

*Proof.* Ad '$\subseteq$': Let $w^* \in \partial f(x^*)$. For every $x \in X$, $h > 0$ there is

$$\langle w^*, x \rangle = \tfrac{1}{h}\langle w^*, (x^* + hx) - x^* \rangle \leq \tfrac{1}{h}[f(x^* + hx) - f(x^*)]$$

and analogously for $h < 0$ sufficiently small so that $x^* + hx \in \operatorname{dom} f$ we have

$$\langle w^*, x \rangle = \tfrac{1}{h}\langle w^*, (x^* + hx) - x^* \rangle \geq \tfrac{1}{h}[f(x^* + hx) - f(x^*)].$$

Consequently, there is

$$\begin{aligned} \langle \mathcal{G}_f(x^*), x \rangle &= \lim_{h \uparrow 0} \tfrac{1}{h}[f(x^* + hx) - f(x^*)] \\ &\leq \langle w^*, x \rangle \leq \lim_{h \downarrow 0} \tfrac{1}{h}[f(x^* + hx) - f(x^*)] = \langle \mathcal{G}_f(x^*), x \rangle \end{aligned}$$

which yields $w^* = \mathcal{G}_f(x^*)$.

Ad '$\supseteq$': For every $x \in X$ and $h \in (0, 1]$ again sufficiently small there is using the convexity of $f$

$$\begin{aligned} \tfrac{1}{h}[f(x^* + h(x - x^*)) - f(x^*)] &= \tfrac{1}{h}[f((1-h)x^* + hx) - f(x^*)] \\ &\leq \tfrac{1}{h}[(1-h)f(x^*) + hf(x) - f(x^*)] = -f(x^*) + f(x) \end{aligned}$$

and thus

$$f(x^*) + \langle \mathcal{G}_f(x^*), x - x^* \rangle = f(x^*) + \lim_{h \downarrow 0} \tfrac{1}{h}[f(x^* + h(x - x^*)) - f(x^*)] \leq f(x).$$

This means immediately that $\mathcal{G}_f(x^*) \in \partial f(x^*)$ by using the definition of the subdifferential. □

# 3 The Elastic Net

In this chapter we apply the theory provided in Chapter 2 to the *elastic net* functional, often abbreviated by elastic net. This means we show that the elastic net has a unique minimiser and how the optimality condition using subdifferentials looks like. Finally, we have a look at further stability properties and at the choice of the parameters and the elastic net.

First of all, we need a formal definition of the elastic net.

**Definition 3.0.1 (elastic net).** *Let $\mathcal{H}$ be a Hilbert space and $\mathcal{D} \in \mathcal{L}(\ell^2, \mathcal{H})$ a linear and continuous operator. For every pair of parameters $(\alpha, \beta) \in \mathbb{P}$ and data $y \in \mathcal{H}$ we define the* **elastic net** *$\Phi_{\alpha,\beta} : \ell^2 \to \mathbb{R}_\infty$ by*

$$\Phi_{\alpha,\beta}(x) := \tfrac{1}{2}\|\mathcal{D}x - y\|_{\mathcal{H}}^2 + \alpha\|x\|_{\ell^1} + \tfrac{1}{2}\beta\|x\|_{\ell^2}^2.$$

The set of all *admissible parameters* is

$$\mathbb{P} := \mathbb{R}_0^+ \times \mathbb{R}^+ := \left\{(\alpha, \beta) \in \mathbb{R}^2 : \alpha \geq 0, \beta > 0\right\}.$$

Even if we do not allow $\beta = 0$ we sometimes use $\Phi_{\alpha,0}$ as an abbreviation for $\frac{1}{2}\|\mathcal{D}x - y\|_{\mathcal{H}}^2 + \alpha\|x\|_{\ell^1}$.

It is crucial to note that the $\ell^1$ norm has to be extended to $\ell^2$ by

$$\|x\|_{\ell^1} := \begin{cases} \sum_{n=1}^{\infty} |x_n|, & \text{if } x \in \ell^1 \\ \infty, & \text{if } x \notin \ell^1 \end{cases}$$

for any $x \in \ell^2$.

Another important fact is that the elastic net can be seen as a stabilised version of an only $\ell^1$ penalised functional. With the scalar product

$$\langle (x_1, y_1), (x_2, y_2) \rangle_{\mathcal{H} \times \ell^2} := \langle x_1, x_2 \rangle_{\mathcal{H}} + \langle y_1, y_2 \rangle_{\ell^2}$$

$\mathcal{H} \times \ell^2$ is also a Hilbert space. Then we can easily rewrite the elastic net functional as

$$\begin{aligned} \Phi_{\alpha,\beta}(x) &= \tfrac{1}{2}\|\mathcal{D}x - y\|_{\mathcal{H}}^2 + \alpha\|x\|_{\ell^1} + \tfrac{1}{2}\beta\|x\|_{\ell^2}^2 \\ &= \tfrac{1}{2}\|(\mathcal{D}x, \sqrt{\beta}x) - (y, 0)\|_{\mathcal{H} \times \ell^2}^2 + \alpha\|x\|_{\ell^1} = \tfrac{1}{2}\|\tilde{\mathcal{D}}x - \tilde{y}\|_{\mathcal{H} \times \ell^2}^2 + \alpha\|x\|_{\ell^1}. \end{aligned}$$

In the last term there is $\tilde{\mathcal{D}}x := (\mathcal{D}x, \sqrt{\beta}x)$ and $\tilde{y} := (y, 0)$. Hence, we can use the whole $\ell^1$ minimisation theory for our elastic net functional. Another benefit of this expression is that it reveals the influence of $\beta$. The new operator is injective in any case, even if $\mathcal{D}$ is not.

For further treatment we have to proof that the elastic net has a unique minimiser for every given data and parameters.

## 3.1. Existence and Uniqueness of the Minimiser

To prove a theorem about the existence of a minimiser and its uniqueness, we first need some basic properties of the elastic net. Since we want to use the direct method, see Chapter 2, we need lower semicontinuity, coercivity, boundedness from below and, for the uniqueness, strict convexity. This is all proven in the following lemmata.

**Lemma 3.1.1 (Bredies and Lorenz [2009, page 5]).** *The elastic net is lower semicontinuous.*

*Proof.* We first prove that $\|\cdot\|_{\ell^1} : \ell^2 \to \mathbb{R}_\infty$ is lower semicontinuous. Let $(x^k)_{k\in\mathbb{N}} \in (\ell^2)^{\mathbb{N}}$ be a convergent sequence with $\lim_{k\to\infty} \|x^k - x\|_{\ell^2} = 0$ and $x \in \ell^2$. We know that strong convergence implies weak convergence and this implies convergence of the components, since the unit sequences $(\delta_n^k)_{n\in\mathbb{N}} \in \ell^2$ for every $k \in \mathbb{N}$. This means that for every $n \in \mathbb{N}$ there holds $\lim_{k\to\infty} x_n^k = x^k$. By Fatou's lemma 2.1.17 we conclude

$$\|x\|_{\ell^1} = \sum_{n=1}^{\infty} |x_n| = \sum_{n=1}^{\infty} |\lim_{k\to\infty} x_n^k| \leq \liminf_{k\to\infty} \sum_{n=1}^{\infty} |x_n^k| = \liminf_{k\to\infty} \|x^k\|_{\ell^1}.$$

Next, the functions $\|\cdot\|_{\ell^2}^2 : \ell^2 \to \mathbb{R}_\infty$ and $\|\mathcal{D}(\cdot) - y\|_{\mathcal{H}}^2 : \ell^2 \to \mathbb{R}_\infty$ are continuous and therefore lower semicontinuous. Consequently, there is

$$\begin{aligned}
\Phi_{\alpha,\beta}(x) &= \tfrac{1}{2}\|\mathcal{D}x - y\|_{\mathcal{H}}^2 + \alpha\|x\|_{\ell^1} + \tfrac{1}{2}\beta\|x\|_{\ell^2}^2 \\
&\leq \tfrac{1}{2}\liminf_{k\to\infty} \|\mathcal{D}x^k - y\|_{\mathcal{H}}^2 + \alpha \liminf_{k\to\infty} \|x^k\|_{\ell^1} + \tfrac{1}{2}\beta \liminf_{k\to\infty} \|x^k\|_{\ell^2}^2 \\
&\leq \liminf_{k\to\infty} \left(\tfrac{1}{2}\|\mathcal{D}x^k - y\|_{\mathcal{H}}^2 + \alpha\|x^k\|_{\ell^1} + \tfrac{1}{2}\beta\|x^k\|_{\ell^2}^2\right) = \liminf_{k\to\infty} \Phi_{\alpha,\beta}(x^k).
\end{aligned}$$

□

**Lemma 3.1.2.** *The elastic net is coercive.*

*Proof.* Let $(x^k)_{k\in\mathbb{N}} \in (\ell^2)^{\mathbb{N}}$ be a sequence with $\lim_{k\to\infty} \|x^k\|_{\ell^2} = \infty$. Then there also holds $\lim_{k\to\infty} \frac{1}{2}\beta\|x^k\|_{\ell^2}^2 = \infty$ and

$$\lim_{k\to\infty} \Phi_{\alpha,\beta}(x^k) \geq \lim_{k\to\infty} \tfrac{1}{2}\beta\|x^k\|_{\ell^2}^2 = \infty,$$

as the other two summands are non-negative. □

**Lemma 3.1.3.** *The elastic net is strictly convex.*

*Proof.* We fulfil this proof in several steps. First, we show the convexity of all terms of the functional and the strict convexity of the third term, i.e. 1. $\|\mathcal{D}(\cdot) - y\|_{\mathcal{H}}^2$ is convex, 2. $\|\cdot\|_{\ell^1}$ is convex and 3. $\|\cdot\|_{\ell^2}^2$ is strictly convex. Lemma 2.1.7 gives us that these properties still hold after the multiplication with their scalars. At the end, Lemma 2.1.6 implies the strict convexity.

Ad 1.: Let $x_1, x_2 \in \ell^2$ and $0 \leq \lambda \leq 1$. Using the abbreviations $\hat{x}_1 := \mathcal{D}x_1 - y$ and $\hat{x}_2 := \mathcal{D}x_2 - y$ we prove the convexity by some straight forward calculations.

$$
\begin{aligned}
& \|\mathcal{D}(\lambda x_1 + (1-\lambda)x_2) - y\|_{\mathcal{H}}^2 - \big(\lambda\|\mathcal{D}x_1 - y\|_{\mathcal{H}}^2 + (1-\lambda)\|\mathcal{D}x_2 - y\|_{\mathcal{H}}^2\big) \\
= \;& \|\lambda\hat{x}_1 + (1-\lambda)\hat{x}_2\|_{\mathcal{H}}^2 - \lambda\|\hat{x}_1\|_{\mathcal{H}}^2 - (1-\lambda)\|\hat{x}_2\|_{\mathcal{H}}^2 \\
= \;& \lambda^2\|\hat{x}_1\|_{\mathcal{H}}^2 + 2\lambda(1-\lambda)\langle\hat{x}_1, \hat{x}_2\rangle_{\mathcal{H}} + (1-\lambda)^2\|\hat{x}_2\|_{\mathcal{H}}^2 - \lambda\|\hat{x}_1\|_{\mathcal{H}}^2 - (1-\lambda)\|\hat{x}_2\|_{\mathcal{H}}^2 \\
= \;& (\lambda^2 - \lambda)\|\hat{x}_1\|_{\mathcal{H}}^2 + 2\lambda(1-\lambda)\langle\hat{x}_1, \hat{x}_2\rangle_{\mathcal{H}} + [(1-\lambda)^2 - (1-\lambda)]\|\hat{x}_2\|_{\mathcal{H}}^2 \\
= \;& \lambda(\lambda-1)\|\hat{x}_1\|_{\mathcal{H}}^2 + 2\lambda(1-\lambda)\langle\hat{x}_1, \hat{x}_2\rangle_{\mathcal{H}} + (1-\lambda)[(1-\lambda) - 1]\|\hat{x}_2\|_{\mathcal{H}}^2 \\
= \;& -\lambda(1-\lambda)\|\hat{x}_1\|_{\mathcal{H}}^2 + 2\lambda(1-\lambda)\langle\hat{x}_1, \hat{x}_2\rangle_{\mathcal{H}} - \lambda(1-\lambda)\|\hat{x}_2\|_{\mathcal{H}}^2 \\
= \;& -\lambda(1-\lambda)(\|\hat{x}_1\|_{\mathcal{H}}^2 - 2\langle\hat{x}_1, \hat{x}_2\rangle_{\mathcal{H}} + \|\hat{x}_2\|_{\mathcal{H}}^2) \\
= \;& -\lambda(1-\lambda)\|\hat{x}_1 - \hat{x}_2\|_{\mathcal{H}}^2 \\
= \;& -\underbrace{\lambda(1-\lambda)}_{\geq 0}\underbrace{\|\mathcal{D}(x_1 - x_2)\|_{\mathcal{H}}^2}_{\geq 0} \leq 0
\end{aligned}
$$

Ad 2.: With the use of the triangle inequality it is easy to see that for $x_1, x_2 \in \ell^1$ and $0 \leq \lambda \leq 1$ there holds

$$\|\lambda x_1 + (1-\lambda)x_2\|_{\ell^1} \leq \|\lambda x_1\|_{\ell^1} + \|(1-\lambda)x_2\|_{\ell^1} = \lambda\|x_1\|_{\ell^1} + (1-\lambda)\|x_2\|_{\ell^1}.$$

If either $x_1 \in \ell^2 \setminus \ell^1$ or $x_2 \in \ell^2 \setminus \ell^1$ the right hand side is $\infty$ and the convexity is rather trivial.

Ad 3.: Analogously to 1., we get for any $x_1, x_2 \in \ell^2, x_1 \neq x_2$ and $0 < \lambda < 1$

$$\|\lambda x_1 + (1-\lambda)x_2\|_{\ell^2}^2 - \big(\lambda\|x_1\|_{\ell^2}^2 + (1-\lambda)\|x_2\|_{\ell^2}^2\big) = -\underbrace{\lambda(1-\lambda)}_{>0}\underbrace{\|x_1 - x_2\|_{\ell^2}^2}_{>0} < 0.$$

□

**Theorem 3.1.4.** *The elastic net has a unique minimiser for every given parameters $(\alpha, \beta) \in \mathbb{P}$ and data $y \in \mathcal{H}$.*

*Proof.* The existence is immediately derived from the direct method, see Theorem 2.2.6. We only have to check the assumptions. $X = \ell^2$ is a Hilbert space and therefore, also a reflexive space. Lemma 3.1.1, 3.1.2 and 3.1.3, state that $\Phi_{\alpha,\beta}$ is lower semicontinuous, coercive and strictly convex. Using now Corollary 2.1.23

we obtain that $\Phi_{\alpha,\beta}$ is weak lower semicontinuous. Obviously $\Phi_{\alpha,\beta}$ is bounded from below by 0, since it is a sum of norms with positive multipliers.

The uniqueness of the minimiser is a simple consequence of the strict convexity of $\Phi_{\alpha,\beta}$. Suppose we have two different minimisers $x_1^*, x_2^* \in \ell^2, x_1^* \neq x_2^*$. Directly by definition of strict convexity and the minising property we have a contradiction.

$$\Phi_{\alpha,\beta}(x_1^*) \leq \Phi_{\alpha,\beta}(\tfrac{1}{2}x_1^* + \tfrac{1}{2}x_2^*) < \tfrac{1}{2}\Phi_{\alpha,\beta}(x_1^*) + \tfrac{1}{2}\Phi_{\alpha,\beta}(x_2^*) \leq \Phi_{\alpha,\beta}(x_1^*) \qquad \lightning$$

□

## 3.2. Subdifferential and Optimality Condition

With the help of subdifferential calculus, the optimality condition reads

$$0 \in \partial\Phi_{\alpha,\beta}(x^*).$$

Our aim is now to calculate the subdifferential of $\Phi_{\alpha,\beta}$ at any $x^* \in \ell^1$. Since $\Phi_{\alpha,\beta}$ is proper and $\operatorname{dom}\Phi_{\alpha,\beta} = \ell^1$, we have for any $x^* \in \ell^2 \setminus \ell^1$ that $\partial\Phi_{\alpha,\beta}(x^*) = \emptyset$, see Remark 2.3.15.

**Lemma 3.2.1.** *The subdifferential of $\|\cdot\|_{\ell^1} : \ell^2 \to \mathbb{R}_\infty$ at any $x^* \in \ell^1$ is given by $\partial\|x^*\|_{\ell^1} = \mathcal{S}^{\mathbb{N}}(x^*) \cap \ell^2$.*

*Proof.* Let $x^* \in \ell^1$ be arbitrary chosen.

Ad '$\subseteq$': Since $\ell^2$ is a Hilbert space, we use the Riesz representation theorem 2.1.9, see Remark 2.3.16, and get that $\partial\|x^*\|_{\ell^1} \subseteq \ell^2$. For $w^* \in \partial\|x^*\|_{\ell^1}$ the subdifferential condition reads

$$\sum_{n=1}^{\infty} |x_n| \geq \sum_{n=1}^{\infty} |x_n^*| + \sum_{n=1}^{\infty} w_n^*(x_n - x_n^*) \quad \forall\, x \in \ell^1.$$

Plugging in the sequences $x \cdot \delta^k := (x_n \cdot \delta_n^k)_{n\in\mathbb{N}} \in \ell^1$ for every $k \in \mathbb{N}$, where $\delta_n^k$ denotes the Kronecker delta, we obtain

$$|x_k| \geq |x_k^*| + w_k^*(x_k - x_k^*),$$

which is already covered by the one dimensional case in Example 2.3.18. Therefore, for every $k \in \mathbb{N}$ there is $w_k^* \in \mathcal{S}^1(x_k^*)$ or equivalent $w^* \in \mathcal{S}^{\mathbb{N}}(x^*)$. Since $w^* \in \ell^2$ we have $w^* \in \mathcal{S}^{\mathbb{N}}(x^*) \cap \ell^2$.

Ad '$\supseteq$': Let us consider now any $w^* \in \mathcal{S}^{\mathbb{N}}(x^*) \cap \ell^2$. Then there holds for any $x \in \ell^1$

$$\begin{aligned}\sum_{n=1}^{\infty} |x_n| &= \sum_{n=1}^{\infty} \operatorname{sign}(x_n)x_n + \sum_{n=1}^{\infty} \underbrace{(\operatorname{sign}(x_n^*) - w_n^*)x_n^*}_{=0} \\ &\geq \sum_{n=1}^{\infty} w_n^* x_n + \sum_{n=1}^{\infty} (\operatorname{sign}(x_n^*) - w_n^*)x_n^* = \sum_{n=1}^{\infty} |x_n^*| + \sum_{n=1}^{\infty} w_n^*(x_n - x_n^*),\end{aligned}$$

which means $w^* \in \partial\|x^*\|_{\ell^1}$. □

**Lemma 3.2.2.** *The subdifferential of the elastic net is given by*

$$\partial\Phi_{\alpha,\beta}(x) = \mathcal{D}'(\mathcal{D}x - y) + \alpha\mathcal{S}^{\mathbb{N}}(x) \cap \ell^2 + \beta x.$$

*Proof.* First of all, we calculate the Gâteaux derivatives of $\frac{1}{2}\|\mathcal{D}(\cdot) - y\|_{\mathcal{H}}^2$ and $\frac{1}{2}\beta\|\cdot\|_{\ell^2}^2$ at any $x \in \ell^2$. From the Examples 2.3.8 and 2.3.9 we obtain the derivatives by plugging in $X = \ell^2$ as

$$\mathcal{G}_{\frac{1}{2}\|\mathcal{D}(\cdot)-y\|_{\mathcal{H}}^2}(x) = \langle \mathcal{D}'(\mathcal{D}x - y), \cdot\rangle_{\ell^2} \quad \text{and} \quad \mathcal{G}_{\frac{1}{2}\beta\|\cdot\|_{\ell^2}^2}(x) = \langle \beta x, \cdot\rangle_{\ell^2}.$$

By using Proposition 2.3.24, we get the corresponding subderivatives, since the interior of the domain of these mappings is the whole $\ell^2$. In addition, we are in a Hilbert space and can identify by Riesz representation theorem 2.1.9 the subdifferentials by $\mathcal{D}'(\mathcal{D}x - y)$ and $\beta x$ respectively.

Secondly, the subdifferential of the $\ell^1$ norm is obtained from Lemma 3.2.1.

To conclude the proof we have to stick all three derivatives together by using Proposition 2.3.23. It can be used since the first two summands are continuous in $\ell^2$. □

**Theorem 3.2.3.** *The optimality condition of the minimiser of the elastic net, which is necessary and sufficient, is given by*

$$-(\mathcal{D}'\mathcal{D} + \beta Id)x + \mathcal{D}'y \in \alpha\mathcal{S}^{\mathbb{N}}(x). \tag{3.2.1}$$

*Proof.* The optimality condition in terms of the subdifferential is

$$0 \in \partial\Phi_{\alpha,\beta}(x) = \mathcal{D}'(\mathcal{D}x - y) + \alpha\mathcal{S}^{\mathbb{N}}(x) \cap \ell^2 + \beta x,$$

see Theorem 2.3.19 and Lemma 3.2.2, which is equivalent to

$$-(\mathcal{D}'\mathcal{D} + \beta Id)x + \mathcal{D}'y \in \alpha\mathcal{S}^{\mathbb{N}}(x) \cap \ell^2.$$

Since $-(\mathcal{D}'\mathcal{D} + \beta Id)x + \mathcal{D}'y \in \ell^2$ for every $x \in \ell^2$ the condition can be simplified to the one which we wanted to prove. □

The optimality condition can be splitted into two different conditions by exploiting the definition of the multivalued sign function.

$$|\mathcal{D}'\mathcal{D}x - \mathcal{D}'y|_n \leq \alpha, \qquad x_n = 0 \tag{3.2.2}$$

$$[-(\mathcal{D}'\mathcal{D} + \beta Id)x + \mathcal{D}'y]_n = \alpha\,\mathrm{sign}(x_n), \qquad x_n \neq 0 \tag{3.2.3}$$

Here we have shortened for a sequence $x$ the absolute value of a component $|x_n|$ by $|x|_n$.

With the notion of sparsity, the second condition shows us an important property of the minimiser.

**Definition 3.2.4 (sparsity).** *A sequence $x \in \mathbb{R}^{\mathbb{N}}$ is called* **sparse** *if there exists $N \in \mathbb{N}$ so that for any $n > N$ there is $x_n = 0$ or equivalent*

$$\|x\|_{\ell^0} := \sum_{n=1}^{\infty} |x_n|^0 = \#\{n \in \mathbb{N} : x_n \neq 0\} < \infty.$$

**Theorem 3.2.5 (Schiffler [2010, page 28]).** *If $\alpha > 0$, then the minimiser of the elastic net is sparse.*

*Proof.* Assume that the minimiser of the elastic net $x^* \in \ell^1$ is not sparse, then there are infinitely many $n \in \mathbb{N}$ so that $x_n^* \neq 0$ and from the optimality condition (3.2.3) we get immediately

$$|-(\mathcal{D}'\mathcal{D} + \beta Id)x^* + \mathcal{D}'y|_n = \alpha$$

for those $n$. But then there is

$$\begin{aligned} \infty &= \|-(\mathcal{D}'\mathcal{D} + \beta Id)x^* + \mathcal{D}'y\|_{\ell^2} \\ &\leq \|\mathcal{D}'\mathcal{D}\|\|x^*\|_{\ell^2} + \beta\|x^*\|_{\ell^2} + \|\mathcal{D}'\|\|y\|_{\mathcal{H}} < \infty. \qquad \lightning \end{aligned}$$

□

## 3.3. The Parameters of the Elastic Net

### 3.3.1. Influence of the Parameters

We have proven that the elastic net has a unique minimiser for any given parameters and data, which we denote by $x_{\alpha,\beta}$, i.e.

$$x_{\alpha,\beta} := \operatorname*{argmin}_{x \in \ell^2} \Phi_{\alpha,\beta}(x).$$

Of course the minimiser depends also on the data $y$, but since we want to focus only on the influence of the parameters we suppress this in the notation.

In many applications it is useful to have an insight about the influences of the parameters. The most important result is summarised in the following theorem.

**Theorem 3.3.1 (stability of the minimiser, Jin et al. [2009, page 2]).** *For every data $y \in \mathcal{H}$ the mapping $\Psi : \mathbb{P} \to \ell^2$, $(\alpha, \beta) \mapsto x_{\alpha,\beta}$ is continuous.*

*Proof.* This proof is based on Jin et al. [2009, page 2], but the second part varies in some important details.

We consider the converging sequence $(\alpha_n, \beta_n)_{n\in\mathbb{N}} \in \mathbb{P}^{\mathbb{N}}$ with $\lim_{n\to\infty}(\alpha_n, \beta_n) = (\alpha, \beta) \in \mathbb{P}$ and denote the corresponding sequence of minimisers by $x^n := x_{\alpha_n,\beta_n}$ and $x^* := x_{\alpha,\beta}$. This proof is devided in two parts. First, we show that w-$\lim_{n\to\infty} x^n =$

$x^*$ and second, $\lim_{n\to\infty} \|x^n\|_{\ell^2} = \|x^*\|_{\ell^2}$. Finally, the use of Lemma 2.1.16 completes the proof.

Ad 1.: Every $x^n$ is the minimiser of the corresponding elastic net and the sequence $(\beta_n)_{n\in\mathbb{N}}$ converges, hence,

$$\tfrac{1}{2}\beta_n \|x^n\|_{\ell^2}^2 \leq \Phi_{\alpha_n,\beta_n}(x^n) \leq \Phi_{\alpha_n,\beta_n}(0) = \tfrac{1}{2}\|y\|_{\ell^2}^2$$

and therefore,

$$\|x^n\|_{\ell^2} \leq \beta_n^{-\frac{1}{2}} \|y\|_{\ell^2} \leq \sup_{n\in\mathbb{N}} \beta_n^{-\frac{1}{2}} \|y\|_{\ell^2} < \infty.$$

Since the sequence $(x^n)_{n\in\mathbb{N}}$ is uniformly bounded in the reflexive space $\ell^2$, Theorem 2.2.1 implies that there exists a weak convergent subsequence $(x^{\eta(n)})_{n\in\mathbb{N}}$. We denote the weak limit by $x^w \in \ell^2$. Similar to the lower semicontinuity, Lemma 3.1.1, we observe

$$\begin{aligned}
\Phi_{\alpha,\beta}(x^w) &\leq \tfrac{1}{2}\liminf_{n\to\infty} \|\mathcal{D}x^{\eta(n)} - y\|_{\mathcal{H}}^2 + \alpha \liminf_{n\to\infty} \|x^{\eta(n)}\|_{\ell^1} + \tfrac{1}{2}\beta \liminf_{n\to\infty} \|x^{\eta(n)}\|_{\ell^2}^2 \\
&\leq \tfrac{1}{2}\liminf_{n\to\infty} \|\mathcal{D}x^{\eta(n)} - y\|_{\mathcal{H}}^2 + \liminf_{n\to\infty} \alpha_{\eta(n)} \|x^{\eta(n)}\|_{\ell^1} + \tfrac{1}{2}\liminf_{n\to\infty} \beta_{\eta(n)} \|x^{\eta(n)}\|_{\ell^2}^2 \\
&\leq \liminf_{n\to\infty} \Phi_{\alpha_{\eta(n)},\beta_{\eta(n)}}(x^{\eta(n)}).
\end{aligned}$$

Using some minimising properties and the estimation above we get

$$\begin{aligned}
\limsup_{n\to\infty} \Phi_{\alpha_{\eta(n)},\beta_{\eta(n)}}(x^{\eta(n)}) &\leq \limsup_{n\to\infty} \Phi_{\alpha_{\eta(n)},\beta_{\eta(n)}}(x^*) \\
&= \Phi_{\alpha,\beta}(x^*) \leq \Phi_{\alpha,\beta}(x^w) \leq \liminf_{n\to\infty} \Phi_{\alpha_{\eta(n)},\beta_{\eta(n)}}(x^{\eta(n)}),
\end{aligned}$$

which implies $\Phi_{\alpha,\beta}(x^w) = \Phi_{\alpha,\beta}(x^*)$. Consequently, $x^w = x^*$ as the elastic net has a unique minimiser. Moreover, every subsequence of the minimising sequence $(x^n)_{n\in\mathbb{N}}$ has a subsequence weakly converging to $x^*$, thus, w-$\lim_{n\to\infty} x^n = x^*$ what we wanted to prove.

Ad 2.: Next, we show that $\lim_{n\to\infty} \|x^n\|_{\ell^2} = \|x^*\|_{\ell^2}$, but since $\|\cdot\|_{\ell^2}$ is weak lower semicontinuous it is sufficient to show that $\limsup_{n\to\infty} \|x^n\|_{\ell^2} \leq \|x^*\|_{\ell^2}$.

Assume this is not true, hence, there exists a constant $c := \limsup_{n\to\infty} \|x^n\|_{\ell^2}^2 > \|x^*\|_{\ell^2}^2$ and a weakly convergent subsequence $(x^{\eta(n)})_{n\in\mathbb{N}}$ so that w-$\lim_{n\to\infty} x^{\eta(n)} = x^*$ and $\lim_{n\to\infty} \|x^{\eta(n)}\|_{\ell^2}^2 = c$. Then there is, using the minimising property and the

weak lower semicontinuity,

$$\begin{aligned}
&\liminf_{n\to\infty} \tfrac{1}{2}\|\mathcal{D}x^{\eta(n)} - y\|_{\mathcal{H}}^2 + \alpha_{\eta(n)}\|x^{\eta(n)}\|_{\ell^1} \\
=\ &\liminf_{n\to\infty} \tfrac{1}{2}\|\mathcal{D}x^{\eta(n)} - y\|_{\mathcal{H}}^2 + \alpha_{\eta(n)}\|x^{\eta(n)}\|_{\ell^1} \\
&\quad + \underbrace{\liminf_{n\to\infty} \tfrac{1}{2}\beta_{\eta(n)}\|x^{\eta(n)}\|_{\ell^2}^2 - \lim_{n\to\infty} \tfrac{1}{2}\beta_{\eta(n)}\|x^{\eta(n)}\|_{\ell^2}^2}_{=0} \\
\leq\ &\liminf_{n\to\infty} \Phi_{\alpha_{\eta(n)},\beta_{\eta(n)}}(x^{\eta(n)}) - \tfrac{1}{2}\beta c \leq \liminf_{n\to\infty} \Phi_{\alpha_{\eta(n)},\beta_{\eta(n)}}(x^*) - \tfrac{1}{2}\beta c \\
=\ &\lim_{n\to\infty} \Phi_{\alpha_{\eta(n)},\beta_{\eta(n)}}(x^*) - \tfrac{1}{2}\beta c = \tfrac{1}{2}\|\mathcal{D}x^* - y\|_{\mathcal{H}}^2 + \alpha\|x^*\|_{\ell^1} + \tfrac{1}{2}\beta\left(\|x^*\|_{\ell^2}^2 - c\right) \\
<\ &\tfrac{1}{2}\|\mathcal{D}x^* - y\|_{\mathcal{H}}^2 + \alpha\|x^*\|_{\ell^1} \\
\leq\ &\liminf_{n\to\infty} \tfrac{1}{2}\|\mathcal{D}x^{\eta(n)} - y\|_{\mathcal{H}}^2 + \alpha\|x^{\eta(n)}\|_{\ell^1} \\
\leq\ &\liminf_{n\to\infty} \tfrac{1}{2}\|\mathcal{D}x^{\eta(n)} - y\|_{\mathcal{H}}^2 + \alpha_{\eta(n)}\|x^{\eta(n)}\|_{\ell^1}. \qquad \lightning
\end{aligned}$$

□

**Remark 3.3.2.** Every minimiser of the elastic net is in $\ell^1$, hence, $\mathrm{Im}(\Psi) \subset \ell^1$.

**Corollary 3.3.3.** *The functions $F_\Phi, F_1, F_2 : \mathbb{P} \to \mathbb{R}$ defined by*

$$F_\Phi(\alpha,\beta) := \Phi_{\alpha,\beta}(x_{\alpha,\beta}), \qquad F_1(\alpha,\beta) := \|x_{\alpha,\beta}\|_{\ell^1} \quad \text{and} \quad F_2(\alpha,\beta) := \|x_{\alpha,\beta}\|_{\ell^2}$$

*are continuous.*

*Proof.* We can rewrite the functions as $F_\Phi = \Phi_{\alpha,\beta}|_{\ell^1} \circ \Psi$, $F_1 = \|\cdot\|_{\ell^1}|_{\ell^1} \circ \Psi$ and $F_2 = \|\cdot\|_{\ell^2}|_{\ell^1} \circ \Psi$, thus, they are all compositions of continuous functions. □

We have considered the cases that both parameters are positive and the one with vanishing $\alpha$. Let us now consider the case that $\beta$ is vanishing. Since $\Phi_{\alpha,0}$ is not strictly convex, we can not expect that the minimiser is unique. Hence, for any $\gamma \geq 0$ we denote the minimiser of $\Phi_{\alpha,0}$, which also minimises $\gamma\|\cdot\|_{\ell^1} + \frac{1}{2}\|\cdot\|_{\ell^2}^2$, by $x_{\alpha,0}^\gamma$.

**Proposition 3.3.4 ($\beta \downarrow 0$, Jin et al. [2009, page 4]).** *Let $(\alpha_n,\beta_n)_{n\in\mathbb{N}} \in \mathbb{P}^{\mathbb{N}}$ be a convergent sequence with $\lim_{n\to\infty}(\alpha_n,\beta_n) = (\alpha,0), \alpha > 0$ and $\lim_{n\to\infty} \frac{\alpha_n-\alpha}{\beta_n} = \gamma \geq 0$. Then there is*

$$\lim_{n\to\infty} x_{\alpha_n,\beta_n} = x_{\alpha,0}^\gamma \quad \text{in } \ell^2.$$

*Proof.* Let us use the notation $x^n := x_{\alpha_n,\beta_n}$ and $x^* := x_{\alpha,0}^\gamma$. We first prove that w-$\lim_{n\to\infty} x^n = x^*$ and secondly, $\lim_{n\to\infty}\|x^n\|_{\ell^1} = \|x^*\|_{\ell^1}$. Finally, from Lemma 2.1.18 we derive

$$\lim_{n\to\infty}\|x^n - x^*\|_{\ell^2} \leq \lim_{n\to\infty}\|x^n - x^*\|_{\ell^1} = 0,$$

which completes the proof.

Ad 1.: Analogously to the proof of Theorem 3.3.1, there exists a subsequence $(x^{\eta(n)})_{n\in\mathbb{N}}$ weakly converging to $x^w$. Since $x^{\eta(n)}, x^*$ are minimisers and $\beta_{\eta(n)} > 0$ there is with $\gamma_n := \frac{\alpha_{\eta(n)}-\alpha}{\beta_{\eta(n)}}$

$$\begin{aligned}\gamma_n\|x^{\eta(n)}\|_{\ell^1} + \tfrac{1}{2}\|x^{\eta(n)}\|_{\ell^2}^2 &= \beta_{\eta(n)}^{-1}[\alpha_{\eta(n)}\|x^{\eta(n)}\|_{\ell^1} + \tfrac{1}{2}\beta_{\eta(n)}\|x^{\eta(n)}\|_{\ell^2}^2 - \alpha\|x^{\eta(n)}\|_{\ell^1}]\\ &= \beta_{\eta(n)}^{-1}[\Phi_{\alpha_{\eta(n)},\beta_{\eta(n)}}(x^{\eta(n)}) - \Phi_{\alpha,0}(x^{\eta(n)})]\\ &\le \beta_{\eta(n)}^{-1}[\Phi_{\alpha_{\eta(n)},\beta_{\eta(n)}}(x^*) - \Phi_{\alpha,0}(x^*)] = \gamma_n\|x^*\|_{\ell^1} + \tfrac{1}{2}\|x^*\|_{\ell^2}^2.\end{aligned}$$

Using these inequation as well as the weak lower semicontinuity and the definition of $\gamma$ leads to

$$\begin{aligned}\gamma\|x^w\|_{\ell^1} + \tfrac{1}{2}\|x^w\|_{\ell^2}^2 &\le \lim_{n\to\infty}\gamma_n \liminf_{n\to\infty}\|x^{\eta(n)}\|_{\ell^1} + \tfrac{1}{2}\liminf_{n\to\infty}\|x^{\eta(n)}\|_{\ell^2}^2\\ &\le \liminf_{n\to\infty}\gamma_n\|x^{\eta(n)}\|_{\ell^1} + \tfrac{1}{2}\|x^{\eta(n)}\|_{\ell^2}^2\\ &\le \liminf_{n\to\infty}\gamma_n\|x^*\|_{\ell^1} + \tfrac{1}{2}\|x^*\|_{\ell^2}^2 = \gamma\|x^*\|_{\ell^1} + \tfrac{1}{2}\|x^*\|_{\ell^2}^2.\end{aligned}$$

Since the minimising element of $\gamma\|\cdot\|_{\ell^1} + \frac{1}{2}\|\cdot\|_{\ell^2}^2$ is unique, we have $x^w = x^*$. With the same arguments as at the proof of Theorem 3.3.1 we conclude w-$\lim_{n\to\infty} x^n = x^*$.

Ad 2.: As at the previous proof, it is sufficient to show that $\limsup_{n\to\infty}\|x^n\|_{\ell^1} \le \|x^*\|_{\ell^1}$. We assume that this is not fulfilled. Therefore, there exists a constant $c := \limsup_{n\to\infty}\|x^n\|_{\ell^1} > \|x^*\|_{\ell^1}$ and subsequence $(x^{\eta(n)})_{n\in\mathbb{N}}$ so that w-$\lim_{n\to\infty} x^{\eta(n)} = x^*$ and a $\lim_{n\to\infty}\|x^{\eta(n)}\|_{\ell^1} = c$. Then we obtain by using the minimising property

$$\begin{aligned}&\liminf_{n\to\infty}\tfrac{1}{2}\|\mathcal{D}x^{\eta(n)} - y\|_{\mathcal{H}}^2\\ =\ &\liminf_{n\to\infty}\tfrac{1}{2}\|\mathcal{D}x^{\eta(n)} - y\|_{\mathcal{H}}^2 + \underbrace{\liminf_{n\to\infty}\alpha_{\eta(n)}\|x^{\eta(n)}\|_{\ell^1} - \lim_{n\to\infty}\alpha_{\eta(n)}\|x^{\eta(n)}\|_{\ell^1}}_{=0}\\ &+ \underbrace{\liminf_{n\to\infty}\tfrac{1}{2}\beta_{\eta(n)}\|x^{\eta(n)}\|_{\ell^2}^2}_{=0}\\ \le\ &\liminf_{n\to\infty}\Phi_{\alpha_{\eta(n)},\beta_{\eta(n)}}(x^{\eta(n)}) - \alpha c \le \liminf_{n\to\infty}\Phi_{\alpha_{\eta(n)},\beta_{\eta(n)}}(x^*) - \alpha c\\ =\ &\lim_{n\to\infty}\Phi_{\alpha_{\eta(n)},\beta_{\eta(n)}}(x^*) - \alpha c = \tfrac{1}{2}\|\mathcal{D}x^* - y\|_{\mathcal{H}}^2 + \alpha\left(\|x^*\|_{\ell^1} - c\right)\\ <\ &\tfrac{1}{2}\|\mathcal{D}x^* - y\|_{\mathcal{H}}^2 \le \liminf_{n\to\infty}\tfrac{1}{2}\|\mathcal{D}x^{\eta(n)} - y\|_{\mathcal{H}}^2. \qquad \lightning\end{aligned}$$

□

**Remark 3.3.5.** The import case that $\alpha_n \equiv \alpha > 0$ and $\beta_n \downarrow 0$ is included in the above proposition. But there are other cases which might be interesting, at least for scientific reasons. These are for example $\alpha_n \uparrow \alpha > 0$, $\alpha_n \downarrow 0$ as well as $\alpha_n \downarrow \alpha > 0$ with $\lim_{n\to\infty}\frac{\alpha_n-\alpha}{\beta_n} = \infty$, e.g. $\alpha_n := \alpha + \frac{1}{n}, \beta_n := \frac{1}{n^2}$. These are not yet considered.

Additionally to the continuity of $F_\Phi$, we can prove further smoothness of this function.

**Theorem 3.3.6 (Jin et al. [2009, page 2]).** *The function $F_\Phi$ is total differentiable in $\mathring{\mathbb{P}}$, especially*

$$\partial_\alpha F_\Phi(\alpha,\beta) = \|x_{\alpha,\beta}\|_{\ell^1} \quad \textit{and} \quad \partial_\beta F_\Phi(\alpha,\beta) = \tfrac{1}{2}\|x_{\alpha,\beta}\|_{\ell^2}^2.$$

*Proof.* We prove the claim in two steps. First, we show that $\partial_\alpha F_\Phi(\alpha,\beta) = \|x_{\alpha,\beta}\|_{\ell^1}$ and second, $\partial_\beta F_\Phi(\alpha,\beta) = \frac{1}{2}\|x_{\alpha,\beta}\|_{\ell^2}^2$. The continuity of the partial derivatives, see Corollary 3.3.3, completes the proof.

Let $(\alpha,\beta) \in \mathring{\mathbb{P}}$ be arbitrary chosen. For $h \in \mathbb{R}$ and small enough so that $\alpha + h \geq 0$ there is by using the minimising property

$$\Phi_{\alpha,\beta}(x_{\alpha+h,\beta}) - \Phi_{\alpha,\beta}(x_{\alpha,\beta}) \geq 0 \quad \text{and} \quad \Phi_{\alpha+h,\beta}(x_{\alpha+h,\beta}) - \Phi_{\alpha+h,\beta}(x_{\alpha,\beta}) \leq 0.$$

Hence, there is with the identity $\Phi_{\alpha+h,\beta}(x) = \Phi_{\alpha,\beta}(x) + h\|x\|_{\ell^1}$

$$\begin{aligned} F_\Phi(\alpha+h,\beta) - F_\Phi(\alpha,\beta) &= \Phi_{\alpha+h,\beta}(x_{\alpha+h,\beta}) - \Phi_{\alpha,\beta}(x_{\alpha,\beta}) \\ &= \Phi_{\alpha,\beta}(x_{\alpha+h,\beta}) - \Phi_{\alpha,\beta}(x_{\alpha,\beta}) + h\|x_{\alpha+h,\beta}\|_{\ell^1} \\ &\geq h\|x_{\alpha+h,\beta}\|_{\ell^1} \end{aligned}$$

and

$$\begin{aligned} F_\Phi(\alpha+h,\beta) - F_\Phi(\alpha,\beta) &= \Phi_{\alpha+h,\beta}(x_{\alpha+h,\beta}) - \Phi_{\alpha,\beta}(x_{\alpha,\beta}) \\ &= \Phi_{\alpha+h,\beta}(x_{\alpha+h,\beta}) - \Phi_{\alpha+h,\beta}(x_{\alpha,\beta}) + h\|x_{\alpha,\beta}\|_{\ell^1} \\ &\leq h\|x_{\alpha,\beta}\|_{\ell^1}. \end{aligned}$$

For $h > 0$ we derive by dividing by $h$

$$\|x_{\alpha+h,\beta}\|_{\ell^1} \leq \tfrac{1}{h}[F_\Phi(\alpha+h,\beta) - F_\Phi(\alpha,\beta)] \leq \|x_{\alpha,\beta}\|_{\ell^1}$$

and respectively for $h < 0$

$$\|x_{\alpha,\beta}\|_{\ell^1} \leq \tfrac{1}{h}[F_\Phi(\alpha+h,\beta) - F_\Phi(\alpha,\beta)] \leq \|x_{\alpha+h,\beta}\|_{\ell^1}.$$

Taking the limit and using Corollary 3.3.3 about the continuity of $\|x_{\alpha,\beta}\|_{\ell^1}$ in $\alpha$ implies $\partial_\alpha F_\Phi(\alpha,\beta) = \|x_{\alpha,\beta}\|_{\ell^1}$.

Similar to the first part there is

$$\begin{aligned} F_\Phi(\alpha,\beta+h) - F_\Phi(\alpha,\beta) &= \Phi_{\alpha,\beta}(x_{\alpha,\beta+h}) - \Phi_{\alpha,\beta}(x_{\alpha,\beta}) + h\tfrac{1}{2}\|x_{\alpha,\beta+h}\|_{\ell^2}^2 \\ &\geq \tfrac{1}{2}h\|x_{\alpha,\beta+h}\|_{\ell^2}^2, \end{aligned}$$

$$\begin{aligned} F_\Phi(\alpha,\beta+h) - F_\Phi(\alpha,\beta) &= \Phi_{\alpha+h,\beta}(x_{\alpha,\beta+h}) - \Phi_{\alpha+h,\beta}(x_{\alpha,\beta}) + h\tfrac{1}{2}\|x_{\alpha,\beta+h}\|_{\ell^2}^2 \\ &\leq \tfrac{1}{2}h\|x_{\alpha,\beta}\|_{\ell^2}^2, \end{aligned}$$

hence, $\partial_\beta F_\Phi(\alpha,\beta) = \frac{1}{2}\|x_{\alpha,\beta}\|_{\ell^2}^2$. □

As for the function $F_\Phi$ we can derive further results for $F_1$ and $F_2$. This is summarised in the following proposition.

**Proposition 3.3.7.** *The function $F_1$ is monotonically decreasing in $\alpha$ and $F_2$ is monotonically decreasing in $\beta$, i.e.*

1. *For every $\beta > 0$ there is* $\alpha_1 \leq \alpha_2 \quad \Rightarrow \quad F_1(\alpha_1, \beta) \geq F_1(\alpha_2, \beta)$ *and*
2. *for every $\alpha \geq 0$ there is* $\beta_1 \leq \beta_2 \quad \Rightarrow \quad F_2(\alpha, \beta_1) \geq F_2(\alpha, \beta_2)$.

*Proof.* We prove only the first part, since the proof ot the second part is analogue. This proof is a simple consequence of an idea, which is already exploited in the previous proof. This is for any $h > 0$ there holds

$$\Phi_{\alpha+h,\beta}(x) = \Phi_{\alpha,\beta}(x) + h\|x\|_{\ell^1}.$$

Then using the minimising property, there is for any $h > 0$

$$\begin{aligned} F_1(\alpha + h, \beta) = \|x_{\alpha+h,\beta}\|_{\ell^1} &= h^{-1}\left[\Phi_{\alpha+h,\beta}(x_{\alpha+h,\beta}) - \Phi_{\alpha,\beta}(x_{\alpha+h,\beta})\right] \\ &\leq h^{-1}\left[\Phi_{\alpha+h,\beta}(x_{\alpha,\beta}) - \Phi_{\alpha,\beta}(x_{\alpha+h,\beta})\right] \\ &\leq h^{-1}\left[\Phi_{\alpha+h,\beta}(x_{\alpha,\beta}) - \Phi_{\alpha,\beta}(x_{\alpha,\beta})\right] = h^{-1}h\|x_{\alpha,\beta}\|_{\ell^1} = F_1(\alpha, \beta). \end{aligned}$$

□

**Remark 3.3.8.** This proposition means that the $\ell^1$ norm of the minimiser monotonically decreases in $\alpha$ and the $\ell^2$ norm of the minimiser monotonically decreases in $\beta$.

**Remark 3.3.9.** For any $\beta > 0$ and upper boundary $\overline{\alpha} > 0$ we can define the function $F_{1,\beta} : [0, \overline{\alpha}] \to \mathbb{R}, \alpha \mapsto F_1(\alpha, \beta)$. As we have seen already, this function is monotonically decreasing and continuous, hence, its derivative exists almost everywhere [Elstrodt, 2004, page 299] and $\partial_\alpha F_{1,\beta} \in L^1(0, \overline{\alpha})$, even if we do not know an explicit expression of it. In the same way this can be obtained for $F_2$.

### 3.3.2. The Choice of the Parameters

The choice of the parameter $\alpha$ and $\beta$ is a rather difficult question, but there are two things what we have to keep in mind when choosing the parameters.

On the one hand, as we have seen above, the influence of the parameters to the solution of the minimising problem is not chaotic, i.e. some mappings are continuous, differentiable or monotonic. On the other hand, we are looking for a solution of a linear equation, thus, we do not want to force the minimiser to be too far away from a real solution if any exists.

Next, we prove that if we choose the parameters too large, the solution tends to zero. But the way how it tends to zero is different for the parameters $\alpha$ and $\beta$.

**Theorem 3.3.10 (Schiffler [2010, page 28]).** *There exists an upper bound $\alpha_{max}$ on the choice of $\alpha$ in the sense that $\alpha \geq \alpha_{max}$ if and only if the minimiser of the elastic net is 0.*

*Proof.* From the optimality condition (3.2.2) we observe that 0 is a minimiser if and only if for all $n \in \mathbb{N}$ there is $|\mathcal{D}'y|_n \leq \alpha$. The smallest $\alpha$ which fulfils this is obviously $\alpha_{max} := \max_{n\in\mathbb{N}} |\mathcal{D}'y|_n$.

The only remaining question is the existence of the maximum. We are looking at the sequence $x^* \in \ell^2$ with $x^*_n := |\mathcal{D}'y|_n$. In the trivial case $x^* = 0$ then $\alpha_{max} = 0$. Otherwise there exists a smallest $\underline{n} \in \mathbb{N}$ so that $x^*_{\underline{n}} > 0$. Since $x^* \in \ell^2$ there exists $\overline{n} \in \mathbb{N}$ so that $x^*_{\underline{n}} > x^*_n$ for all $n > \overline{n}$, hence, $\alpha_{max} = \max_{\underline{n}\leq n\leq\overline{n}} |\mathcal{D}'y|_n$. But here we consider only finitely many values and thus $\alpha_{max}$ exists. □

The set of all admissible or rather senseful parameters is bounded in the direction of $\alpha$, as we have proven above. In the direction of $\beta$ we are only able to prove that the solution tends to zero, but without any sharp boundary.

**Lemma 3.3.11.** *Let $(\alpha, \beta_n)_{n\in\mathbb{N}} \in \mathbb{P}^{\mathbb{N}}$ be a sequence of parameters with $\lim_{n\to\infty} \beta_n = \infty$. Then the sequence of minimisers $(x_{\alpha,\beta_n})_{n\in\mathbb{N}} \in (\ell^2)^{\mathbb{N}}$ converges to the zero solution in $\ell^2$.*

*Proof.* This can be proven easily in the following way. When we denote the minimiser of $\Phi_{\alpha,\beta_n}$ by $x^n$, we have for any $n \in \mathbb{N}$

$$\tfrac{1}{2}\beta_n \|x^n\|^2_{\ell^2} \leq \Phi_{\alpha,\beta_n}(x^n) \leq \Phi_{\alpha,\beta_n}(0) = \tfrac{1}{2}\|y\|^2_{\ell^2}$$

and thus, $\|x^n\|^2_{\ell^2} \leq \beta_n^{-1}\|y\|^2_{\ell^2}$, which leads to

$$\lim_{n\to\infty} \|x^n\|^2_{\ell^2} \leq \lim_{n\to\infty} \beta_n^{-1}\|y\|^2_{\ell^2} = 0.$$

□

To summarise the results and to give an overview of all admissible parameters this set is shown in Figure 3.1.

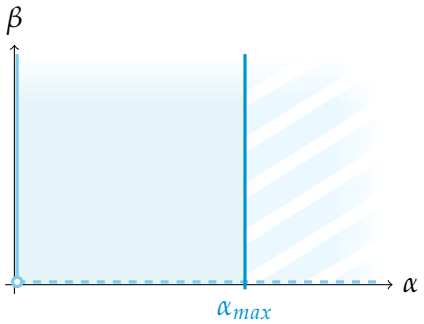



Figure 3.1: The set of all admissible parameters. On the right hand side there is a boundary, since every solution with parameters inside the marked area, i.e. $\alpha$ is larger than $\alpha_{max}$, is the zero solution. In the direction of $\beta$ the solution also tends to zero but without any sharp boundary.

# Regularised Feature Sign Search

In this chapter we propose an algorithm to solve the elastic net. When treating the elastic net as a $\ell^1$ penalised functional there are plenty of algorithms which solve this problem. The *Feature Sign Search (FSS)*, introduced by Lee et al. [2007], showed very often a good performance. But there are problems with rank deficient matrices, as for example when the matrix has more columns than rows, which is the case in our application. In such cases, the FSS may try to invert a singular matrix.

The algorithm we would like to use is called *Regularised Feature Sign Search (RFSS)*. It is a regularised, i.e. the inversion is stabilised, version of the FSS. It was first proposed by Jin et al. [2009]. To prove a result on convergency, we can only consider finite dimensional problems, i.e. $\mathcal{D} : \mathbb{R}^N \to \mathbb{R}^M$ and $y \in \mathbb{R}^M$. But since real examples, i.e. when using a computer, are always finite, this is no drawback for this algorithm.

To state this algorithm we need the notion of *consistency* and have to use an auxiliary functional which is differentiable. These are introduced in the following section.

## 4.1. Consistency

For simplicity of the formulas some new notations are needed. We denote the set of all possible indices of $x \in \mathbb{R}^N$, which is $\{1, \dots, N\}$, by $\mathcal{N}$. For every set $\Gamma \subset \mathcal{N}$ the complement $\mathcal{N} \setminus \Gamma$ is written as $\Gamma_c$. The *restriction* of a column vector $x = (x_n)_{n \in \mathcal{N}}$ on the *active set* $\Gamma \subset \mathcal{N}$ is defined component-wise or row-wise by $x_\Gamma := (x_n)_{n \in \Gamma}$. Analogously, we define the restriction of a matrix column-wise, i.e. for $\mathcal{D} = (d_1, \dots, d_N) = (d_n)_{n \in \mathcal{N}}$, where $d_n \in \mathbb{R}^M$ denotes the columns of the matrix $\mathcal{D}$, there is $\mathcal{D}_\Gamma := (d_n)_{n \in \Gamma}$. Lately we denote the *residual* by $R(x) := \mathcal{D}x - y$.

**Definition 4.1.1 (consistency).** *Let $\Gamma \subset \mathcal{N}, x \in \mathbb{R}^N$ and $\theta \in \{-1, 0, 1\}^N$. The triple $(\Gamma, x, \theta)$ is called* **consistent** *if $x_{\Gamma_c} = \theta_{\Gamma_c} = 0$ and for every $n \in \Gamma$ there is*

$$\operatorname{sign}(x_n) = \theta_n \neq 0.$$

With the notion of consistency we can rewrite the already splitted optimality conditions.

**Proposition 4.1.2 (Optimality Condition and Consistency).** *Let $(\Gamma, x, \theta)$ be a consistent triple. Then $x$ is the minimiser of the elastic net if and only if*

$$\max_{n \in \Gamma_c} |\langle d_n, R(x) \rangle| \leq \alpha \tag{O1}$$

$$\text{and} \qquad x_\Gamma = (\mathcal{D}_\Gamma' \mathcal{D}_\Gamma + \beta Id)^{-1} (\mathcal{D}_\Gamma' y - \alpha \theta_\Gamma). \tag{O2}$$

*Proof.* The optimality conditions (3.2.2) and (3.2.3) are

$$|\mathcal{D}'\mathcal{D}x - \mathcal{D}'y|_n \leq \alpha, \qquad \text{if } x_n = 0 \tag{4.1.1}$$

$$\text{and} \quad [-(\mathcal{D}'\mathcal{D} + \beta Id)x + \mathcal{D}'y]_n = \alpha \operatorname{sign}(x_n), \qquad \text{if } x_n \neq 0. \tag{4.1.2}$$

From the definition of consistency we obtain that $x_n \neq 0$ if and only if $n \in \Gamma$. Thus, for any $n \in \Gamma_c$ there is by using (4.1.1)

$$\alpha \geq |\mathcal{D}'\mathcal{D}x - \mathcal{D}'y|_n = |\mathcal{D}'(\mathcal{D}x - y)|_n = |\mathcal{D}'R(x)|_n.$$

By definition of matrix vector multiplication and of the Euclidean scalar product there is for any $n \in \Gamma_c$

$$\alpha \geq |\mathcal{D}'R(x)|_n = |\langle d_n, R(x) \rangle|,$$

which is fulfilled if and only if (O1) holds.

It is rather trivial to obtain the second optimality condition. Since the active set $\Gamma$ is the set of the indices corresponding to all non-zero components, we can write (4.1.2) as an equation of vectors as

$$-(\mathcal{D}_\Gamma' \mathcal{D}_\Gamma + \beta Id) x_\Gamma + \mathcal{D}_\Gamma' y = \alpha \theta_\Gamma.$$

For any $\beta > 0$ the matrix $\mathcal{D}_\Gamma' \mathcal{D}_\Gamma + \beta Id$ is invertible, thus, this equation is equivalent to (O2). □

**Definition 4.1.3 (auxiliary functional).** *For any $\Gamma \subset \mathcal{N}$ and $\theta \in \{-1, 0, 1\}^N$ we define the* **auxiliary functional** $\Xi_{\theta,\Gamma} : \mathbb{R}^{|\Gamma|} \to \mathbb{R}$ *by*

$$\Xi_{\theta,\Gamma}(z) := \tfrac{1}{2} \|\mathcal{D}_\Gamma z - y\|_2^2 + \alpha \langle \theta_\Gamma, z \rangle + \tfrac{1}{2} \beta \|z\|_2^2.$$

**Remark 4.1.4.** The auxiliary functional is related to the elastic net by

$$\Xi_{\theta,\Gamma}(x_\Gamma) \leq \Phi_{\alpha,\beta}(x)$$

for any triple $(\Gamma, x, \theta)$. Equality holds if and only if the triple is consistent.

The auxiliary functional is differentiable in $z$ for any fixed $\theta$ and $\Gamma$. In addition, it is strictly convex as a sum of two convex and one strictly convex function, see Lemma 2.1.6. Combining these, the unique solution of the minimisation problem $\min_{z \in \mathbb{R}^{|\Gamma|}} \Xi_{\theta,\Gamma}(z)$ is given by the solution of

$$\nabla_\Gamma \Xi_{\theta,\Gamma}(z) := (\partial_{z_n} \Xi_{\theta,\Gamma}(z))_{n \in \Gamma} = 0.$$

**Lemma 4.1.5.** *Let $(\Gamma, x, \theta)$ be a triple. Then $x$ fulfils the optimality condition* (O2) *if and only if $\nabla_\Gamma \Xi_{\theta,\Gamma}(x_\Gamma) = 0$.*

*Proof.* For every $n^* \in \Gamma$ there is, using $\partial_{n^*}(\mathcal{D}_\Gamma x_\Gamma)_m = \partial_{n^*} \sum_{n\in\Gamma}(d_n)_m x_n = (d_{n^*})_m$,

$$\begin{aligned} 0 &= \partial_{n^*}\Xi_{\theta,\Gamma}(x_\Gamma) \\ &= \partial_{n^*}\tfrac{1}{2}\sum_{m=1}^{M}(\mathcal{D}_\Gamma x_\Gamma - y)_m^2 + \partial_{n^*}\alpha\sum_{n\in\Gamma}\theta_n x_n + \partial_{n^*}\tfrac{1}{2}\beta\sum_{n\in\Gamma}x_n^2 \\ &= \sum_{m=1}^{M}(d_{n^*})_m(\mathcal{D}_\Gamma x_\Gamma - y)_m + \alpha\theta_{n^*} + \beta x_{n^*} \\ &= \langle d_{n^*}, \mathcal{D}_\Gamma x_\Gamma - y\rangle + \alpha\theta_{n^*} + \beta x_{n^*} \\ &= \langle d_{n^*}, R(x)\rangle + \alpha\theta_{n^*} + \beta x_{n^*}. \end{aligned} \tag{4.1.3}$$

This can also be written in a vector equation as

$$0 = \nabla_\Gamma \Xi_{\theta,\Gamma}(x_\Gamma) = \mathcal{D}_\Gamma'(\mathcal{D}_\Gamma x_\Gamma - y) + \alpha\theta_\Gamma + \beta x_\Gamma$$

and solved by $x_\Gamma = (\mathcal{D}_\Gamma'\mathcal{D}_\Gamma + \beta Id)^{-1}(\mathcal{D}_\Gamma' y - \alpha\theta_\Gamma)$. □

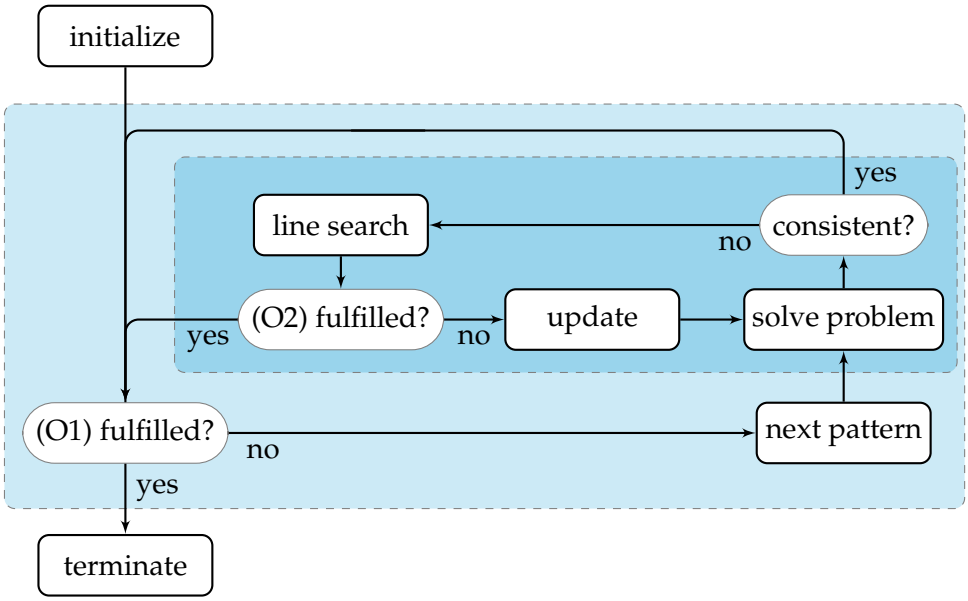
## 4.2. The Algorithm

The main idea of the algorithm is the following. According to Proposition 4.1.2, we want to find a consistent triple so that the optimality conditions (O1) and (O2) are fulfilled. But in both of these conditions there hides a problem. First, how large do we have to choose the active set so that (O1) is fulfilled? Second, the right hand side of (O2) depends on the sign of the solution whereas the left hand side is the solution itself. This algorithm tries to guess how the active set looks like and how the signs of the solution are. In other words this algorithm is looking for the signs of the features, which are needed for the solution.

The algorithm with all the details is provided in Table 4.1. Additionally, we discribe the algorithm roughly. The guesses of the active set and the sign vector are done iteratively. Since we are looking for a sparse solution, i.e. with lots of zero components, our first guess is by default that the active set is empty. If this was not the correct solution we have to increase our active set. We increase the active set by the index, which violates the optimality condition (O1) most. In addition, we guess what the sign of the solution at our new index is. This guess looks rather arbitrary but, as we will prove in Lemma 4.3.3, it is the correct guess. Next, we solve the optimality condition (O2) on the active set, which depends on our chosen active set and sign vector. Since for a sparse solution the active set is very small, the system of linear equations we need to solve is small as well.

In the whole algorithm it is very important that the current triple is consistent. Hence, before we have a look if our calculated solution is the solution of the elastic net we have to check if the solution is consistent or not. If this is not the case it

Table 4.1.: The algorithm of the RFSS. '$\leftarrow$' means that the variable on the left hand side gets the value of the right hand side.




| | |
|---|---|
| **initialise** | Set $(\Gamma, x, \theta)^0 \leftarrow (\emptyset, 0, 0)$ and $k \leftarrow 0$. |
| **next pattern** | Increase $k \leftarrow k+1$. We take the index which is violating (O1) the most, i.e. fits best to the residual, $$n^* \in \underset{n \in \Gamma_c^k}{\operatorname{argmax}} \lvert\langle d_n, R(x^{k-1})\rangle\rvert$$ and update the active set $\Gamma^k \leftarrow \Gamma^{k-1} \cup \{n^*\}$. Then the sign vector is updated by $\theta_n^k \leftarrow \theta_n^{k-1}, n \neq n^*$ and $$\theta_{n^*}^k \leftarrow -\operatorname{sign}\langle d_{n^*}, R(x^{k-1})\rangle.$$ |
| **solve problem** | Solve the minimisation problem on the active set by $x_{\Gamma_c^k}^k \leftarrow 0$ and $$x_{\Gamma^k}^k \leftarrow \left(\mathcal{D}'_{\Gamma^k}\mathcal{D}_{\Gamma^k} + \beta Id\right)^{-1}\left(\mathcal{D}'_{\Gamma^k} y - \alpha\theta_{\Gamma^k}^k\right).$$ |
| **line search** | Increase $k \leftarrow k+1$. Find the smallest $\lambda \in (0,1]$ so that for $x^k \leftarrow x^{k-2} + \lambda(x^{k-1} - x^{k-2})$ there exists an index $m^*$ with $$\operatorname{sign}(x_{m^*}^k) \neq \operatorname{sign}(x_{m^*}^{k-2}).$$ Remove the index from the active set $\Gamma^k \leftarrow \Gamma^{k-1} \setminus \{m^*\}$ and update the sign vector by $\theta_n^k \leftarrow \theta_n^{k-1}, n \neq m^*$ and $\theta_{m^*}^k \leftarrow 0$. |
| **update** | Increase $k \leftarrow k+1$ and update the active set by $\Gamma^k \leftarrow \Gamma^{k-1}$ as well as the sign vector by $\theta^k \leftarrow \theta^{k-1}$. |

has to be fixed. Otherwise, we can iteratively continue with the already described procedure.

Overall it is not easy to see why this algorithm really does what it has to and that it converges to the solution of the elastic net. In fact, it is not obvious that it converges at all. This is proven in the next section.

**Remark 4.2.1.** Any other consistent triple would also be possible for initialisation if we only modify the entrance behaviour slightly.

## 4.3. Proof of Convergency

Before we can start proving the convergency of the algorithm, we have to check that the algorithm is well-defined. There are plenty of assumptions that need to be fulfilled so that the next step is useful. In the following, we show that these assumptions are indeed fulfilled if the starting triple is the trivial one $(\emptyset, 0, 0)$. As mentionend above any other consistent starting triple is also allowed when we extend the entrance behaviour of the algorithm slightly. Here we only proof the well-definedness for the trivial starting triple.

The abbreviations $\Xi^k := \Xi_{\theta^k,\Gamma^k}$ and $(\Gamma, x, \theta)^k := (\Gamma^k, x^k, \theta^k)$ help us to guarantee readability and to shorten the expressions.

**Lemma 4.3.1.** *The inner loop guarantees that* (O2) *is fulfilled.*

*Proof.* There are two ways out of the inner loop. First, we leave directly after 'solve problem' but then (O2) is fulfilled, since it was calculated according to solve this equation. The second way out of the inner loop is when (O2) is fulfilled after the 'line search'. □

**Corollary 4.3.2.** *At the beginning of any iteration of the outer loop, i.e. before 'next pattern', the optimality condition* (O2) *is fulfilled.*

*Proof.* The initial triple fulfils trivially the optimal condition (O2) since $\Gamma^0$ is empty. After any iteration the condition is also fulfilled, see Lemma 4.3.1. □

**Lemma 4.3.3 (Schiffler, 2010, page 47).** *We have predicted* $\operatorname{sign}(x^k_{n^*})$ *in 'next pattern' correctly, i.e.* $\theta^k_{n^*} := -\operatorname{sign}\langle d_{n^*}, R(x^{k-1})\rangle = \operatorname{sign}(x^k_{n^*})$.

*Proof.* Let $(\Gamma, x, \theta)^k$ denote the triple after 'solve problem' and $(\Gamma, x, \theta)^{k-1}$ the one before choosing 'next pattern'. Since $x^k$ is optimal for $\Xi^k$ there is $\Xi^k(x^k) \leq \Xi^k(x^{k-1})$ and as a result of the convexity there is for every $0 \leq h \leq 1$

$$\Xi^k(x^{k-1} + h[x^k - x^{k-1}]) \leq h\Xi^k(x^k) + (1-h)\Xi^k(x^{k-1}) \leq \Xi^k(x^{k-1}). \tag{4.3.1}$$

We also know that $|\langle d_{n^*}, R^{k-1}\rangle| > \alpha$ and $x^{k-1}_{n^*} = 0$. Using equation (4.1.3) leads

to

$$\begin{aligned}\operatorname{sign}(\partial_{n^*}\Xi^k(x^{k-1})) &= \operatorname{sign}(\langle d_{n^*}, R(x^{k-1})\rangle + \alpha\theta^k_{n^*} + \beta x^{k-1}_{n^*}) \\ &= \operatorname{sign}(\langle d_{n^*}, R(x^{k-1})\rangle + \alpha\theta^k_{n^*}) \\ &= \operatorname{sign}\langle d_{n^*}, R(x^{k-1})\rangle = -\theta^k_{n^*},\end{aligned} \tag{4.3.2}$$

because this is how $\theta^k_{n^*}$ is chosen .

Since $\theta^k_{\Gamma^{k-1}} = \theta^{k-1}_{\Gamma^{k-1}}$ and $x^{k-1}$ is optimal for $\Xi^{k-1}$, see Corollary 4.3.2, there is

$$\nabla_{\Gamma^{k-1}}\Xi^k(x^{k-1}) = \nabla_{\Gamma^{k-1}}\Xi^{k-1}(x^{k-1}) = 0. \tag{4.3.3}$$

Assume now that we have predicted the sign wrong, i.e. $\operatorname{sign}(x^k_{n^*}) \neq \theta^k_{n^*}$ or $\theta^k_{n^*} \cdot x^k_{n^*} < 0$. Using the Taylor series expansion with a sufficiently small stepsize $h$ we obtain

$$\begin{aligned}0 &\overset{(4.3.1)}{\geq} \Xi^k(x^{k-1} + h[x^k - x^{k-1}]) - \Xi^k(x^{k-1}) \\ &= \langle \nabla_{\Gamma^k}\Xi^k(x^{k-1}), x^{k-1} + h[x^k - x^{k-1}] - x^{k-1}\rangle + \mathcal{O}(h^2) \\ &= h\langle \nabla_{\Gamma^k}\Xi^k(x^{k-1}), x^k - x^{k-1}\rangle + \mathcal{O}(h^2) \\ &\overset{(4.3.3)}{=} h\,\partial_{n^*}\Xi^k(x^{k-1})x^k_{n^*} + \mathcal{O}(h^2) \\ &\overset{(4.3.2)}{=} -\underbrace{h|\partial_{n^*}\Xi^k(x^{k-1})|}_{>0}\underbrace{\theta^k_{n^*}x^k_{n^*}}_{<0} + \mathcal{O}(h^2) > 0. \qquad \lightning\end{aligned}$$

**Lemma 4.3.4.** *The 'line search' always finds a change of the sign if the triple of the previous iteration, either inner or outer loop, was consistent.*

*Proof.* The current setting is the following. The triple $(\Gamma, x, \theta)^{k-1}$, obtained from the previous iteration of the inner or outer loop, is by assumption consistent. We also know that $(\Gamma, x, \theta)^k$ is inconsistent since we have not left the inner loop.

We have to consider two different cases. First, this is the first iteration of the inner loop and second, it is not.

Ad 1.: We know that $\Gamma^k = \Gamma^{k-1} \cup \{n^*\}$, $\theta^k_n = \theta^{k-1}_n$ for $n \in \Gamma^{n-1}$ and $\operatorname{sign}(x^k_{n^*}) = \theta^k_{n^*}$, see Lemma 4.3.3. Then there exists an index $n \in \Gamma^{n-1}$ so that

$$\operatorname{sign}(x^k_n) \neq \theta^k_n = \theta^{k-1}_n = \operatorname{sign}(x^{k-1}_n).$$

Ad 2.: In this case, there is $\Gamma^k = \Gamma^{k-1}$, $\theta^k = \theta^{k-1}$ and hence, there is $n \in \Gamma^{n-1}$ so that

$$\operatorname{sign}(x^k_n) \neq \theta^k_n = \theta^{k-1}_n = \operatorname{sign}(x^{k-1}_n).$$

In any case there exists a change of the sign. □

**Proposition 4.3.5.** *The inner loop guarantees that the resulting triple is consistent.*

*Proof.* The resulting triple of the inner loop $(\Gamma, x, \theta)^k$ which might not be consistent is that which is made by 'line search'.

Since the initial triple is consistent we prove this by induction. Assume that the triple of the previous iteration $(\Gamma, x, \theta)^{k-2}$ is consistent.

We have to differ two situations. First, this is the first iteration of the inner loop and second, it is not.

Ad 1.: We know that the new index $n^*$ is consistent, i.e. $\theta_{n^*}^{k-1} = \operatorname{sign}(x_{n^*}^{k-1})$. By definition of $(\Gamma, x, \theta)^k$ we have for any $n \in \Gamma^k \setminus \{n^*\}$ that $\operatorname{sign}(x_n^k) = \operatorname{sign}(x_n^{k-2}) = \theta_n^{k-2} = \theta_n^k$. Moreover, we have proven that $\operatorname{sign}(x_{n^*}^k) = \theta_{n^*}^k$, see Lemma 4.3.3. Overall, we have that for any $n \in \Gamma^k$ there is $\operatorname{sign}(x_n^k) = \theta_n^k$. Furthermore, $x_{m^*}^k = 0$ and $\theta_{m^*}^k = 0$, which correspond to the removed index $m^*$, hence, the triple is consistent.

Ad 2.: This case is analogue to the previous one. The only difference is that there is no new index $n^*$. □

**Theorem 4.3.6 (Schiffler [2010, page 46]).** *The function $\Phi_{\alpha,\beta}(x^k)$ is strictly decreasing in k, i.e. every iteration of the inner loop strictly reduces the value of the functional.*

*Proof.* We know that the resulting triple of the last inner iteration $(\Gamma, x, \theta)^{k-1}$ is consistent, see Proposition 4.3.5. If this is the first iteration of the inner loop there is $\Gamma^k = \Gamma^{k-1} \cup \{n^*\}$, $\theta_{\Gamma^{k-1}}^k = \theta_{\Gamma^{k-1}}^{k-1}$ and $x_{n^*}^{k-1} = 0$, hence,

$$\begin{aligned}
\Phi_{\alpha,\beta}(x^{k-1}) &= \Xi^{k-1}(x^{k-1}) \\
&= \tfrac{1}{2}\| \sum_{n\in\Gamma^{k-1}} d_n x_n^{k-1} - y\|_2^2 + \alpha \sum_{n\in\Gamma^{k-1}} \theta_n^{k-1} x_n^{k-1} + \tfrac{1}{2}\beta \sum_{n\in\Gamma^{k-1}} |x_n^{k-1}|^2 \\
&= \tfrac{1}{2}\| \sum_{n\in\Gamma^{k}} d_n x_n^{k-1} - y\|_2^2 + \alpha \sum_{n\in\Gamma^{k}} \theta_n^{k-1} x_n^{k-1} + \tfrac{1}{2}\beta \sum_{n\in\Gamma^{k}} |x_n^{k-1}|^2 \\
&= \Xi^k(x^{k-1}).
\end{aligned}$$

In any other iteration there is $\Gamma^k = \Gamma^{k-1}$, $\theta^k = \theta^{k-1}$ and immediately $\Phi_{\alpha,\beta}(x^{k-1}) = \Xi^{k-1}(x^{k-1}) = \Xi^k(x^{k-1})$. Note that in any case we have

$$\Phi_{\alpha,\beta}(x^{k-1}) = \Xi^k(x^{k-1}). \tag{4.3.4}$$

Additionally, we know that $(\Gamma^k, x^{k-1}, \theta^k)$ violates (O2) but $(\Gamma^k, x^k, \theta^k)$ does not or in other words $x^k$ is optimal for $\Xi^k$ and $x^{k-1}$ is not, thus,

$$\Xi^k(x^k) < \Xi^k(x^{k-1}). \tag{4.3.5}$$

Next, we consider two different cases. First, the triple after 'solve problem' is consistent and we terminate the inner loop and second, it is not.

Ad 1.: If the resulting triple is consistent we have

$$\Phi_{\alpha,\beta}(x^k) = \Xi^k(x^k) \overset{(4.3.5)}{<} \Xi^k(x^{k-1}) \overset{(4.3.4)}{=} \Phi_{\alpha,\beta}(x^{k-1}).$$

Ad 2.: In this case we have done a line search so that $(\Gamma, x, \theta)^{k+1}$ is consistent. By the convexity of $\Xi^k$ and $x_{m^*}^{k+1} = 0$ there is

$$\begin{aligned}\Phi_{\alpha,\beta}(x^{k+1}) &= \Xi^{k+1}(x^{k+1}) = \Xi^k(x^{k+1}) = \Xi^k(\lambda x^k + (1-\lambda)x^{k-1}) \\ &\leq \lambda\Xi^k(x^k) + (1-\lambda)\Xi^k(x^{k-1}) \\ &\overset{(4.3.5)}{<} \lambda\Xi^k(x^{k-1}) + (1-\lambda)\Xi^k(x^{k-1}) = \Xi^k(x^{k-1}) \overset{(4.3.4)}{=} \Phi_{\alpha,\beta}(x^{k-1}).\end{aligned}$$

□

**Theorem 4.3.7 (Schiffler [2010, page 44]).** *The RFSS converges globally to the unique minimiser of the elastic net in finitely many steps.*

*Proof.* We know from Theorem 4.3.6 that every iteration of the inner loop strictly decreases the value of $\Phi_{\alpha,\beta}(x^k)$. Additionally, we know that the next iterate $x^{k+1}$ depends only on the active set, hence an active set does not occur twice. Especially there are no loops. The RFSS converges in finitely many steps, since there are only finitely many possibilities for the active set (in fact there are $2^N$ possibilities). □

# Analysis of Sea Floor Pressure Data

This chapter is devoted to the application of sparsity in the analysis of sea floor pressure data. In Section 5.1, we have a look at how sparsity by means of the elastic net can be used to analyse these data sets. In addition, four other methods, namely Harmonic Decomposition, Wavelet Decomposition, EMD and EEMD, are applied to these data sets as well. These methods are described in Section 5.2. Finally, the results of all five methods are presented in Section 5.3.

## 5.1. Sparse Decomposition

Sparse Decomposition contains several steps to obtain a solution to the decomposition problem. First, one has to find a proper *dictionary* which depends a lot on the application. Next, $\ell^1$ minimisation by RFSS is used to select which pattern in the dictionary are needed for a suitable decomposition. At last, we perform $\ell^2$ minimisation with the chosen pattern to get the best fitting solution, i.e. minimising

$$\|\mathcal{D}_\Gamma x_\Gamma - y\|_2^2 + \beta\|x_\Gamma\|_2^2,$$

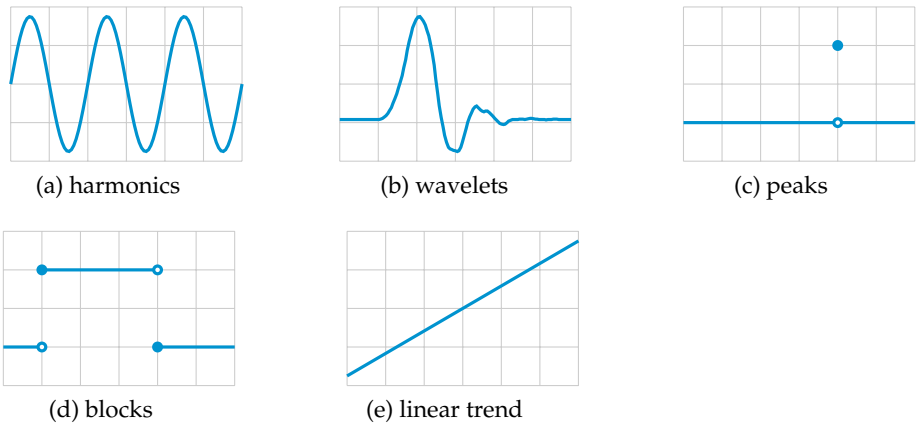
by $x_\Gamma = (\mathcal{D}_\Gamma'\mathcal{D}_\Gamma + \beta Id)^{-1}\mathcal{D}_\Gamma' y$ with a small $\beta > 0$ to ensure the invertibility of the matrix.

Let us have a closer look at the choices needed for Sparse Decomposition. These are the choice of the dictionary and of the parameters, in particular the sparsity parameter $\alpha$.

### 5.1.1. Dictionary

The choice of the dictionary is an important step to analyse sea floor pressure data. Every effect we want to have in our decomposition has to be in the dictionary as a single pattern or as a superposition of patterns which represents this effect.

We have chosen harmonics and wavelets of a large scale to represent the tides. Short time effects, like earthquakes or errors in the measurement, are represented very well by wavelets of a small scale and by peaks. The long time feature or trend detection is done with blocks of constant pressure and a global linear trend because these pattern do neither have a high resolution in time to represent short time features nor do they represent tidal effects. The chosen pattern are shown in Figure 5.1.



(a) harmonics (b) wavelets (c) peaks

(d) blocks (e) linear trend

Figure 5.1.: The dictionary used by Sparse Decomposition contains harmonics and wavelets for tidal constituents, wavelets and peaks for short time components as well as blocks and a linear trend for the long time components.

First of all, we have a closer look at the harmonic components. The Fourier basis is the natural choice to represent the tides. If we want that our dictionary is similar to the tidal components, we also have to use translations of those harmonics. The Fourier basis has the size of the number of measurements in the data set. Our data sets contain up to 120,000 measurements, which means that we would have 120,000 harmonics in our dictionary and we did not consider translations yet. Since we can not store $120,000 \times 120,000$ matrices, which are not sparse, we have to get rid of the unnecessary harmonic components by use of *extra knowledge* about physical oceanography [Stewart, 1997].

Table 5.1.: Fundamental Tidal Frequencies with their periods and sources [Stewart, 1997].

| | Frequency [1/day] | Period | Source |
|---|---|---|---|
| $\omega_1$ | 9.661e-1 | 1 lunar day | Local mean lunar time |
| $\omega_2$ | 3.660e-2 | 1 month | Moon's mean longitude |
| $\omega_3$ | 2.738e-3 | 1 year | Sun's mean longitude |
| $\omega_4$ | 3.095e-4 | 9 years | Longitude of moon's perigee |
| $\omega_5$ | -1.471e-4 | 19 years | Longitude of moon's ascending node |
| $\omega_6$ | 1.307e-7 | 20,940 years | Longitude of sun's perigee |

Doodson [1922] has claimed, that there exist *fundamental frequencies* so that the frequency of the *principal tidal constituents* may be written as $\omega = \sum_{n=1}^{6} \lambda_n \omega_n$, with fundamental frequencies $\omega_n$ according to Table 5.1 and *Doodson numbers* $\lambda_n$ to Ta-

ble 5.2. These fundamental frequencies are due to the movement of the moon and the sun with respect to the earth. In most applications some of the fundamental frequencies can be dropped since the tidal prediction using the remaining fundamental frequencies is accurate enough.

Table 5.2.: Principal Tidal Constituents with their Doodson numbers, frequency and period [Stewart, 1997].

| | | Doodson numbers | | | | | | Frequency | Period |
|---|---|---|---|---|---|---|---|---|---|
| | | $\lambda_1$ | $\lambda_2$ | $\lambda_3$ | $\lambda_4$ | $\lambda_5$ | $\lambda_6$ | [1/day] | [days] |
| Semi-diurnal $\lambda_1 = 2$ | Principal lunar | 2 | 0 | 0 | 0 | 0 | 0 | 1.932e-0 | 5.175e-1 |
| | Principal solar | 2 | 2 | -2 | 0 | 0 | 0 | 2.000e-0 | 5.000e-1 |
| | Lunar elliptic | 2 | -1 | 0 | 1 | 0 | 0 | 1.896e-0 | 5.274e-1 |
| | Lunisolar | 2 | 2 | 0 | 0 | 0 | 0 | 2.005e-0 | 4.986e-1 |
| Diurnal $\lambda_1 = 1$ | Lunisolar | 1 | 1 | 0 | 0 | 0 | 0 | 1.003e-0 | 9.973e-1 |
| | Principal lunar | 1 | -1 | 0 | 0 | 0 | 0 | 9.295e-1 | 1.076e-0 |
| | Principal solar | 1 | 1 | -2 | 0 | 0 | 0 | 9.973e-1 | 1.003e-0 |
| | Elliptic lunar | 1 | -2 | 0 | 1 | 0 | 0 | 8.932e-1 | 1.120e-0 |
| Long Period $\lambda_1 = 0$ | Fortnightly | 0 | 2 | 0 | 0 | 0 | 0 | 7.320e-2 | 1.366e+1 |
| | Monthly | 0 | 1 | 0 | -1 | 0 | 0 | 3.629e-2 | 2.755e+1 |
| | Semiannual | 0 | 0 | 2 | 0 | 0 | 0 | 5.476e-3 | 1.826e+2 |

By using this knowledge, we can reduce the number of harmonic pattern from a full basis with all its translations to only these eleven harmonics and its translations. In most applications 30 translations for each frequency are adequate, hence, there are 330 harmonics in the dictionary.

The wavelet basis we use is created by translations and scalings of the wavelet *Daubechies 5* and its scaling function up to level four. The support of this wavelet is $2 \cdot 5 = 10$ measurements, which is sufficiently small to represent most short time features. The coarse levels of this basis can also represent tidal effects.

The next pattern we want to discuss are the peak functions. These are in most cases not similar to effects we want to decompose but nevertheless very useful. The peaks represent failures by the measuring device or by transferring data. If the dictionary does not include peak functions the RFSS will try to fix these failures by choosing lots of other patterns which is clearly not physically senseful. Moreover, if the sampling interval is not short enough to have more than one measurement during an earthquake, which lasts usually between 10 to 30 minutes, these earthquakes appear as peak functions in the data sets. As seen in Section 1.1 the sampling interval of our data sets are up to 60 minutes, thus, we might observe earthquakes as peaks in these data sets.

The long time features are represented by blocks and a global linear trend in our

dictionary. These blocks are chosen to have a length, depending on the duration of the data set, from one week to five years. Of course a drift of tectonic plates does not occur like a block function but these functions represent lots of trends in a sufficient manner. The simplicity and the bounded support of these functions help us that we do not tend to overstate the physical meaning of the computed trend.

At the last point of this section, we want to discuss the computed dictionary operator for the four used data sets. Some important facts are given in Table 5.3. For all data sets the dictionary operator was created using Matlab®'s parallel toolbox with two Quad-Core AMD Opteron(tm) Processor 2376, 2.3 GHz per core, and in total 15.7 GB RAM. Most of the pattern, in fact twice the number of measurements, are wavelets, scaling functions and peaks. Their short support is indispensable for us to store such huge matrices. In Chapter 3 we have shown that the zero minimiser is obtained if the sparsity parameter $\alpha$ is chosen larger than $\alpha_{max}$ depending on the operator and the data. These upper bound for a proper choice of $\alpha$ is also shown in the table. But we do not want to compute the zero solution, thus, $\alpha$ is chosen a lot smaller than these prior computed upper bounds.

Table 5.3.: Analysis of the dictionary operator, saved as sparse matrix in Matlab®. *: offset for starting Matlab®'s parallel toolbox included

| | constructing time [s]* | size | memory [MB] | sparsity [%] | $\alpha_{max}$ |
|---|---|---|---|---|---|
| SYN | 16.0 | $1,000 \times 2,397$ | 3.7 | 86.7 | 208.7 |
| MAR | 21.9 | $22,319 \times 45,110$ | 105.9 | 99.1 | 396.3 |
| CORK1 | 202.5 | $117,012 \times 234,437$ | 516.5 | 99.8 | 2,172.5 |
| CORK2 | 109.1 | $78,983 \times 159,011$ | 457.1 | 99.7 | 1,662.7 |

### 5.1.2. Parameters

Another possibility to influence the Sparse Decomposition, next to the choice of the dictionary operator, is the choice of the parameters $\alpha$ and $\beta$.

The choice of the sparsity parameter $\alpha$ is quite difficult. As we mentionend above we have an upper bound $\alpha_{max}$ for a proper choice. In practice, we obtained decompositions, which use not too many pattern and differ pointwise not more than around 0.2 $kPa$ from the data, by choosing $\alpha \in [0.1, 0.25]$ depending on the data and the dictionary operator.

The RFSS needs the invertibility of $\mathcal{D}_\Gamma' \mathcal{D}_\Gamma$ in every iteration. If the size of the active set is less or equal to the number of measurements this is equivalent to the full rank of $\mathcal{D}_\Gamma$. Otherwise, if the size of the active set is too large this operator is never invertible. Using the parameter $\beta$ we invert $\mathcal{D}_\Gamma' \mathcal{D}_\Gamma + \beta Id$ instead which is

invertible for any $\beta > 0$, because all eigenvalues of this matrix are greater than or equal to $\beta$. To guarantee a stable inversion of the matrix, $\beta$ should be chosen a lot greater than the numerical zero which is about 1e-16. Also $\beta$ should not be chosen too large so that $\mathcal{D}_\Gamma'\mathcal{D}_\Gamma + \beta Id \approx \mathcal{D}_\Gamma'\mathcal{D}_\Gamma$, hence, a good choice might be $\beta =$ 1e-10.

## 5.2. Other Tools for Decomposition

In this section we present four other methods to analyse sea floor pressure data sets. First, we discuss methods, classical in time series analysis, namely Harmonic and Wavelet Decomposition. They are based on the Fourier and wavelet transform, respectively. In addition, we also present two novel methods in time series analysis. These are the Empirical Mode Decomposition and its enhancement Ensemble Empirical Mode Decomposition.

All methods have in common that they seek for a new representation of the data set as a superposition of several pattern. The classical approaches achieve this by basis transformation to the Fourier or wavelet domain. In contrast the novel approaches calculate a dynamic basis depending on the given data set.

### 5.2.1. Harmonic Decomposition

The Harmonic Decomposition is mainly based on the discrete Fourier transform and the decomposition is done in the frequency domain. First of all, the time series is transformed by the discrete Fourier transform. By low-pass, band-pass and high-pass filtering, we can decompose it to components, which frequencies correspond to physical effects. The inverse transform yields a decomposed signal, which components differ in frequency.

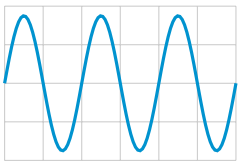


Figure 5.2.: The imaginary part of the Harmonic basis function $s^3$.

To be more precise, let us denote a given time series by $y = (y_n)_{n=0,\dots,N-1} \in \mathbb{R}^N$. The Harmonic basis of $\mathbb{C}^N$ is given by $(s^k)_{k=0,\dots,N-1}$ as

$$s_n^k = \exp(\tfrac{2\pi i}{N}kn) = \cos(\tfrac{2\pi}{N}kn) + i\sin(\tfrac{2\pi}{N}kn),$$

using Euler's formula. To illustrate this basis the imaginary part of $s^3$ is shown in Figure 5.2. Note that the complex conjugate of $s_n^k$ is given by $\exp(-\frac{2\pi i}{N}kn)$. Hence, the discrete Fourier transform of $y$ is $\hat{y} \in \mathbb{C}^N$ defined by

$$\hat{y}_k := \langle y, s^k\rangle_{\mathbb{C}^N} = \sum_{n=0}^{N-1} y_n \exp(-\tfrac{2\pi i}{N}kn).$$

The components are also known as the *Fourier coefficients*. The basis function $s^k$ is associated to the frequency $k$ for $0 \le k \le \frac{N}{2}$ and to $N-k$ if $\frac{N}{2} < k \le N-1$. The

first claim is rather obvious. To show the second one we only have a look at the real part since the imaginary part is a time shift of the real part. Let us suppose that $\frac{N}{2} < k \leq N-1$ and define $\tilde{k} := N - k$. Then using a trigonometric addition formula, we obtain

$$\begin{aligned}\cos(\tfrac{2\pi}{N}kn) &= \cos(\tfrac{2\pi}{N}(N-\tilde{k})n) = \cos(2\pi n - \tfrac{2\pi}{N}\tilde{k}n)\\ &= \cos(2\pi n)\cos(-\tfrac{2\pi}{N}\hat{k}n) - \sin(2\pi n)\sin(-\tfrac{2\pi}{N}\tilde{k}n)\\ &= \cos(-\tfrac{2\pi}{N}\tilde{k}n) = \cos(\tfrac{2\pi}{N}\tilde{k}n),\end{aligned}$$

which means that the associated frequency is $\tilde{k} = N - k$.

After choosing the frequencies, which correspond to the components we want to obtain, we can decompose the signal by filtering the Fourier coefficients and applying the inverse discrete Fourier transform given by

$$y_n = \frac{1}{N}\sum_{k=0}^{N-1} \hat{y}_k \exp(\tfrac{2\pi i}{N}kn).$$

The above vectors can be seen as functions on the interval $[0,1]$. In our application the data sets can be seen as functions on the interval $[0,T]$, where $T$ is the duration of the time series. Hence, the $k$th component of the discrete Fourier transform corresponds to the frequency $\frac{k}{T}$ for $0 \leq k \leq \frac{N}{2}$ and to $\frac{N-k}{T}$ for $\frac{N}{2} < k \leq N-1$. By plugging in the duration of the data set MAR, we obtain for instance that the 30th Fourier coefficient corresponds to the frequency $1/day$.

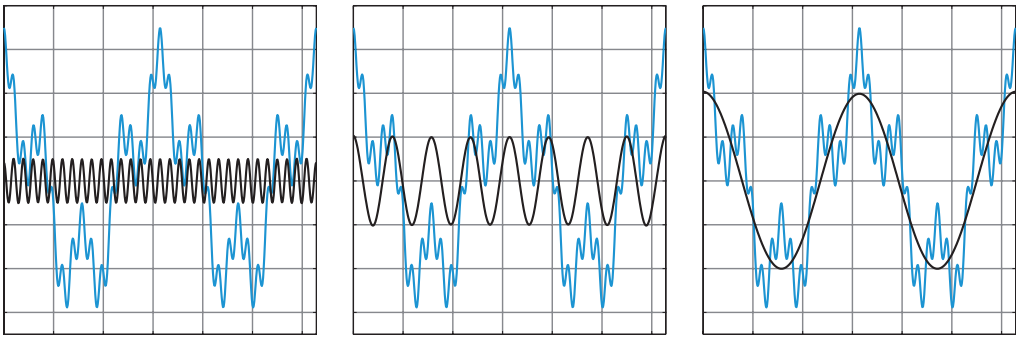


Figure 5.3.: Example of Harmonic decomposition. The time series which is a superposition of harmonics is shown in blue and the decomposition in black from left to right with increasing frequencies.

For this application we have chosen the cut-off frequencies around $0.2/day$ and $4/day$. This means that the Harmonic basis functions with frequencies up to $0.2/day$ are associated with the long time component, the basis functions with a frequency between $0.2/day$ and $4/day$ are associated with the medium time component and the short time component has frequencies higher than $4/day$. In this

way we know that the major tidal features are in the medium time component since their frequencies are around $1/day$ and $2/day$.

To illustrate the Harmonic decomposition an example decomposition is shown in Figure 5.3. This example time series consists of three harmonics which are well detected and separated.

For further information we refer to Gröchenig [2001] and Stark [2005].

### 5.2.2. Wavelet Decomposition

The second method we want to discuss is the Wavelet Decomposition. This method is very similar to the Harmonic Decomposition except that the basis used for transformation is a wavelet basis. The transformed signal can be decomposed by choosing which scales corresponds to which component. Again, the inverse transformation yields to the decomposed data set. But let us again be more precise about what is going on.

To stress the main notion of the wavelet transformation let us consider a data set $y \in L^2(\mathbb{R},\mathbb{R})$. For any chosen orthogonal *mother wavelet* $\psi$ we get the corresponding orthonormal basis for $L^2(\mathbb{R},\mathbb{R})$ given by

$$\{\psi_{m,k} \in L^2(\mathbb{R},\mathbb{R}) : \psi_{m,k}(x) = 2^{-m/2}\psi(2^{-m/2}x - k), m,k \in \mathbb{Z}\}.$$

It is important to note that the basis functions $\psi_{m,k}$ are shifted and scaled version of the mother wavelet $\psi$. Then the signal can be rewritten as

$$y = \sum_{m,k\in\mathbb{Z}} \langle y, \psi_{m,k}\rangle_{L^2(\mathbb{R},\mathbb{R})}\psi_{m,k}.$$

Hence, we can decompose the data set into components of different scale.

In practice we do not consider the function space $L^2(\mathbb{R},\mathbb{R})$ but only $\mathbb{R}^N$ and a discrete version of the transformation is needed. The discrete wavelet transformation is given by a vector of detail and approximation coefficients. Iteratively up to a predefined coarse level, the transform yields coefficients which represent the details and the basic features of the data set. In the every further iteration these coefficients are associated to coarser features. This is exploited by the Wavelet Decomposition.

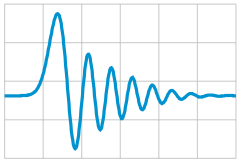


Figure 5.4.: Wavelet Daubechies 40 (db40, scaling function)

In the application we have to predefine the mother wavelet and the coarsest level of the decomposition. One reasonable choice for the coarsest level is $\lfloor \log_2(N) \rfloor$, where $N$ is the number of measurements of the data set and $\lfloor\cdot\rfloor$ is the floor function, i.e. $\lfloor x \rfloor := \max\{m \in \mathbb{Z} : m \leq x\}$. The reason for this choice is that the number of approximation coefficients at level $n$ is roughly $N/2^n$, thus, this is the 'maximal' decomposition. The mother wavelet is chosen as one of *Daubechies* wavelets

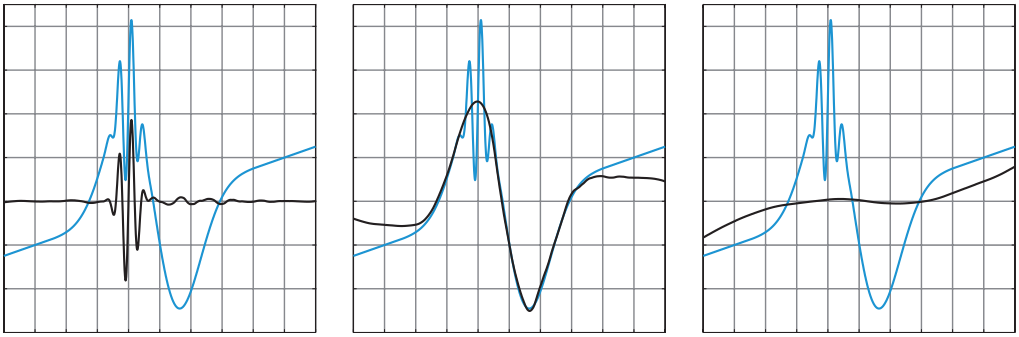


Figure 5.5.: Example using the Wavelet Decomposition. The time series is shown in blue and the decomposition in black from left to right with an increasing scale. The data is a superposition of an affine trend and two harmonics, which are dumped by a gaussian.

since they have a similar shape like the tides. Best results were obtained by using Daubechies 40 as the mother wavelet, whose scaling function is illustrated in Figure 5.4.

An example decomposition of a signal by using wavelets is given in Figure 5.5.

Further information about wavelets and complete information about the discrete wavelet transform can be found either in a theoretical manner in Louis et al. [1997] or in an applied manner with lots of examples in Stark [2005].

### 5.2.3. Empirical Mode Decomposition

The third method we would like to introduce is the Empirical Mode Decomposition (EMD) invented by Huang et al. [1998]. The classical methods use a predefined and fixed basis for transformation as we have seen already. Contrary, the EMD can be seen as computing a new basis depending on the data set. The EMD computes for any given time signal finitely many *Intrinsic Mode Functions* (IMF), which are functions satisfying the following two conditions.

1. The number of extrema and the number of zero crossings must differ at most by one.
2. The mean value of the upper and lower envelope is zero at any time, where the upper envelope is computed by interpolating the maxima and the lower envelope by interpolating the minima.

The task of the EMD is to find these IMFs and is done by the *sifting process*.

1. Find all extrema of the given data.
2. If the number of extrema is one or less, we have found all IMFs and terminate the algorithm.

3. Compute the upper and lower envelope by interpolation the maxima and minima and the mean of these.
4. If the mean value is zero, we have found an IMF. Start the sifting process again with the data subtracted by this IMF. Otherwise, start the sifting process again with the data subtracted by the mean.

This is also shown in Figure 5.6.

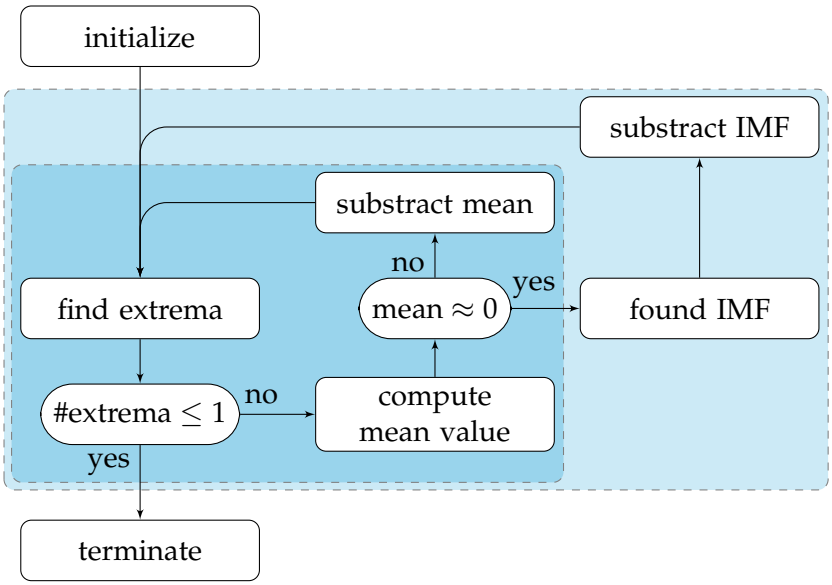



Figure 5.6.: The algorithm of the EMD. The coloured areas represent the loops of the algorithm.

As we have seen we have done no a priori choices. In fact one can choose an interpolating scheme but this is predefined by cubic splines which works very well as tested with some example time series. Since subtracting an IMF reduces the number of extrema, this algorithm terminates for every finite signal. A disadvantage of this method is that it is empirical and has no solid theoretical foundation. Compared to the Harmonic and Wavelet Decomposition, which are mainly based on the fast Fourier transform and the discrete wavelet transform, the computing effort to decompose the data by EMD is higher but still tolerable. A detailed introduction to the EMD is given in Huang et al. [1998].

In Figure 5.7 we see the EMD applied to a time series, which is a super position of harmonics and a linear trend.

### 5.2.4. Ensemble Empirical Mode Decomposition

A challenge in decomposing time series is to prevent *mode mixing*. Phenomena of similar time scale should be in the same mode (here: IMF) and vice versa phenom-

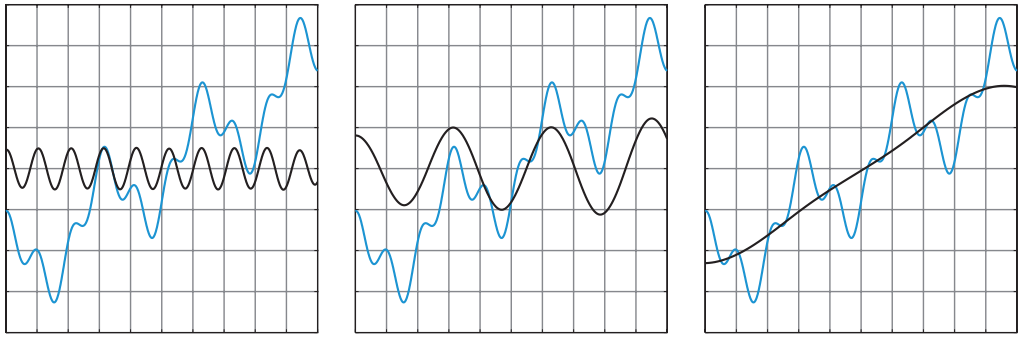
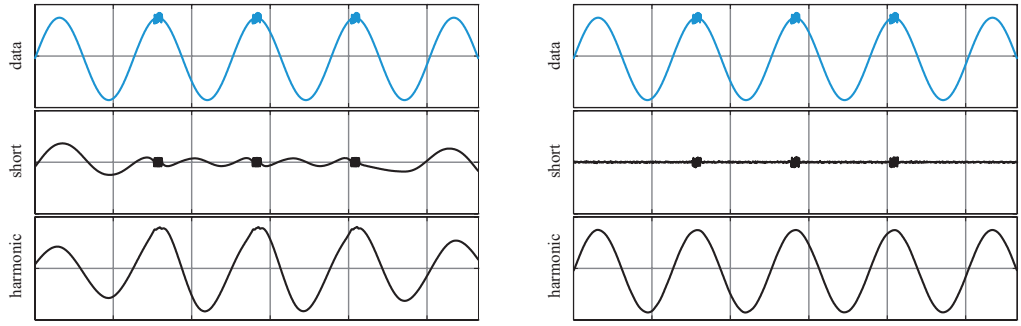


Figure 5.7.: Example of the EMD. The time series is shown in blue and the decomposition in black. On the left and the middle the separated component is harmonic like. The right image shows the separated trend. The data set is a superposition of two harmonics and a linear trend




(a) Decomposition by EMD

(b) Decomposition by EEMD

Figure 5.8.: Example of the EEMD. The usual EMD fails to decompose the time series because of mode mixing. Contrary, the noise assisted EEMD performs very well and separates the high frequency component from the low frequent harmonic. The EEMD has taken 50 trials with a white noise of standard deviation ratio of 0.1.

ena of a different scale are expected to be in a different mode. For this purpose Wu and Huang suggested in 2009 a noise assisted data analysis method, called EEMD [Wu and Huang, 2009]. For any given signal, white noise is added and then decomposed using the EMD. At the end - after numerous trials - we take the mean of the IMFs. 'By adding finite noise, the EEMD elimated largely the mode mixing problem and preserve physical uniqueness of decomposition. Therefore, the EEMD represents a major improvement of the EMD method.' [Wu and Huang, 2009]. Of course this is more time consuming than the usual EMD and we have to check if this is worthwhile in our application or not.

As parameters we have to specify the ratio of the standard deviation of the

signal and the standard deviation of the added white noise and the number of trials. The standard deviation ratio in our application varies between 0.1 and 0.2 and the number of iterations varies between 50 and 100.

To see the ability of the EEMD to prevent mode mixing when EMD fails we have a look at Figure 5.8. The data set is a harmonic with a low frequency and on top of the maxima there are high frequent harmonics. This example was also proposed by Wu and Huang [2009].

## 5.3. Results

In this section we present and discuss the results. To emphasise the variation of the pressure we have preprocessed the data by subtracting the mean value. The mean value of the data sets is about 25,000 $kPa$ and the features we are looking for might have an amplitude of only a few $kPa$, hence, subtracting the mean is crucial for plotting these data sets.

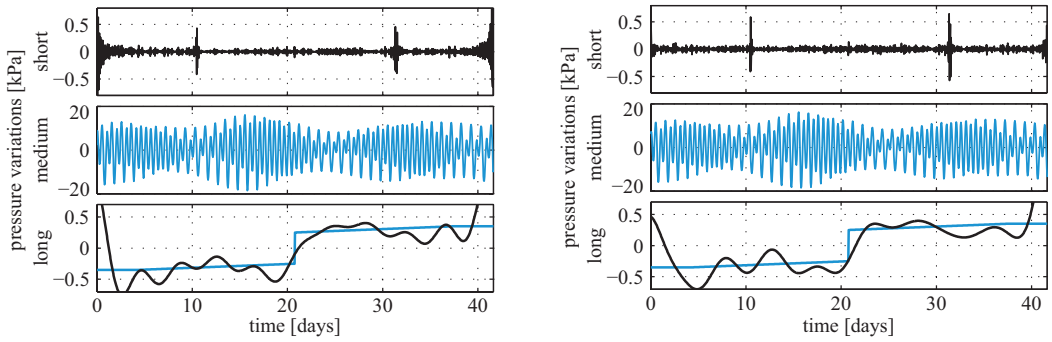
For the Harmonic and Wavelet Decomposition we have used Matlab®'s functions `fft` and `wavedec`, respectively. These are implemented very fast and fulfil all the requirements for the Harmonic and Wavelet Decomposition. The algorithms for EMD and EEMD are implemented by Rilling [2007] and Wu [downloaded 04/2011]. These are implemented as a m-file and thus, do not compete with `fft` and `wavedec` in time. This is also true for the RFSS, which was implemented by Schiffler [2009]. All three algorithms can be found on the world wide web, see our references.

All algorithms extract the tides very well, hence, this is not mentioned in further discussions.

### 5.3.1. Decomposition of SYN

First, we have a look at the decomposition of the data SYN, which is shown in Figure 5.9. Additionally, the short and long time components for this data set are presented in Figure 5.10. This data set is a synthetic one and therefore the calculated decompositions can be compared to the real one. The minimal goal for all methods is to decompose the data set into three main components, which are 'short', 'medium' and 'long' and represent features with periods or wavelength shorter than the tides, the tides themselves and the one with longer lasting features. As we see the methods' ability for extracting all five components varies substantially. The Sparse Decomposition is able to decompose the short time features very well but fail by separating the step from the ramp. All other decomposition methods are also not able to separate these two components from each other. The Harmonic and Wavelet Decomposition were capable to separate the short time features from the tides but can not separate them from the noise. The coarsest decomposition is given by EMD and EEMD. These methods are only able to decompose the medium and long time component.

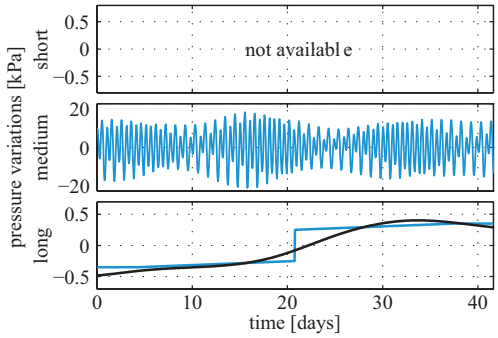
Harmonic, Wavelet and Sparse Decomposition correctly detect the short time



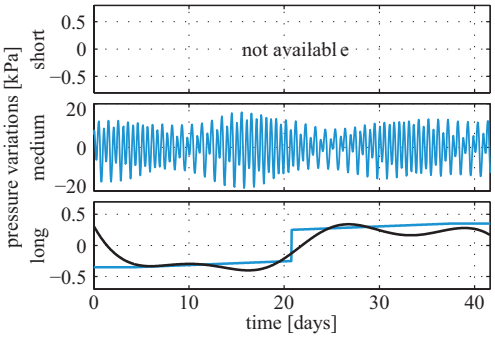

(a) Harmonic Decomposition



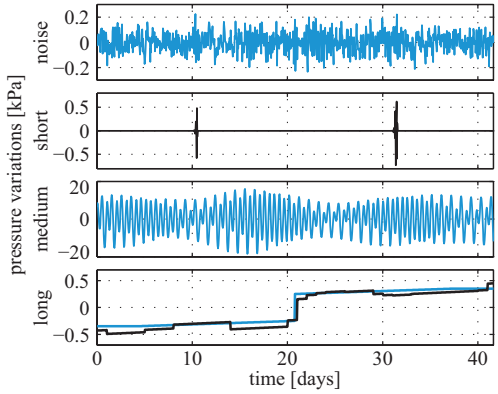

(b) Wavelet Decomposition



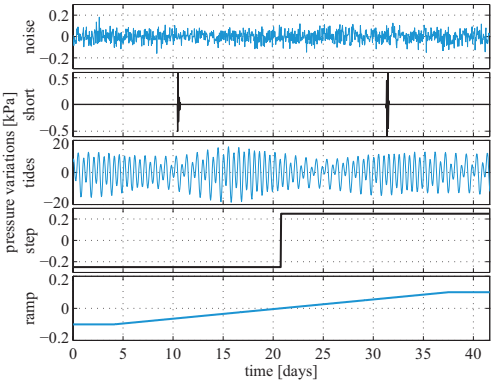

(c) EMD




(d) EEMD

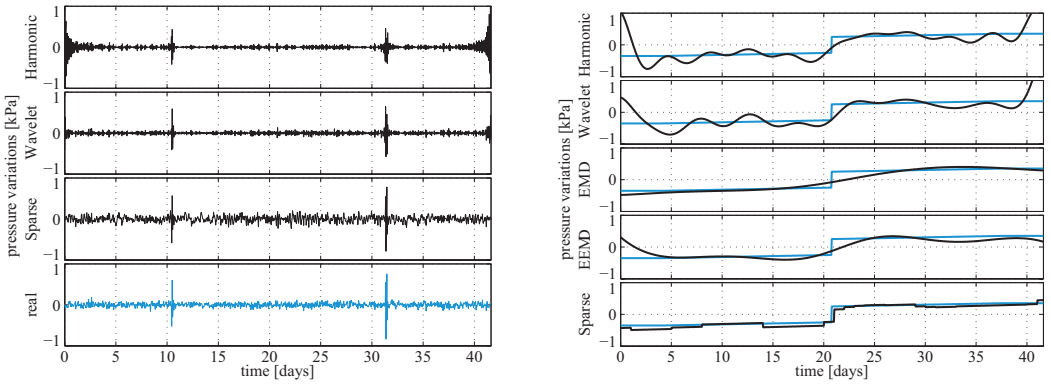



(e) Sparse Decomposition




(f) real decomposition

Figure 5.9.: Decomposition of data set SYN. There are shown from the top to the bottom the short, medium and long time components. The real decomposition can be seen in Figure 5.9(f). EMD and EEMD were not capable to detect features with a higher frequency than the tides.




(a) Short time components of SYN

(b) Long time components of SYN

Figure 5.10.: Left: Short time components of the data set SYN decomposed by the methods Harmonic, Wavelet and Sparse. EMD and EEMD were not capable to decompose features with a higher frequency than the tides. Right: Long time components of the data set SYN decomposed by all five methods. The long time component obtained by the methods is drawn black and the real trend is drawn blue.

features in time but the classical methods tend to underestimate the amplitude. The results of Harmonic and Wavelet Decomposition show very large boundary effects due to filtering, periodisation and the lack of pattern, which fit to all components. Methodically, Sparse Decomposition has no boundary effects when the dictionary is chosen well enough. On the one hand, Sparse Decomposition is not based on filtering nor does it assume a periodic signal. On the other hand, if the dictionary contains local and global pattern, features on one scale do not have to be fitted by pattern of a different scale. Overall, we see that Sparse Decomposition performs best in decomposing the short time features.

The long time features are decomposed very well by all algorithms. Except of Sparse Decomposition the methods are not capable to reproduce the edge of the step. EMD and Sparse Decomposition perform really good in detecting the ramp. Again the results of Harmonic and Wavelet Decomposition are degraded by boundary effects. In summary, all methods reproduce the real long time component quite well, especially the trend of the Sparse Decomposition is very similar to the real one.

Since this is a synthetic data set we do not only compare the results visually but check their quality also in terms of the relative errors, i.e. the quotion of the norm of the error and the norm of the component we are looking for. These results are shown in Table 5.4.

The short time features are decomposed best in terms of relative errors by Wavelet Decomposition in the noisy case. The boundary effects at the short time com-

Table 5.4.: Relative errors of the methods for SYN. The relative error for a real component $c$ and the calculated component $\tilde{c}$ is $\frac{\|c-\tilde{c}\|_2}{\|c\|_2}$. *: no short time component obtained. **: the noise was not seperated from the short time component.

| | | Harmonic | Wavelet | EMD | EEMD | Sparse |
|---|---|---|---|---|---|---|
| relative errors | short + noise | 2.6665 | 0.7007 | -* | -* | 0.9132 |
| | short | -** | -** | -* | -* | 0.3305 |
| | tides | 0.0426 | 0.0381 | 0.0151 | 0.0279 | 0.0133 |
| | ramp + step | 0.8522 | 0.9250 | 0.3026 | 0.4650 | 0.2650 |

ponent of Harmonic Decomposition are very large, thus the relative error is worst of those methods, which were capable to find short time features. Only Sparse Decomposition is able to separate a noise-free short time component which is also very well in terms of relative errors.

The tides are decomposed satisfactorily by all methods as mentioned above. Best performance is shown by EMD and Sparse Decomposition but all the others are really good as well.

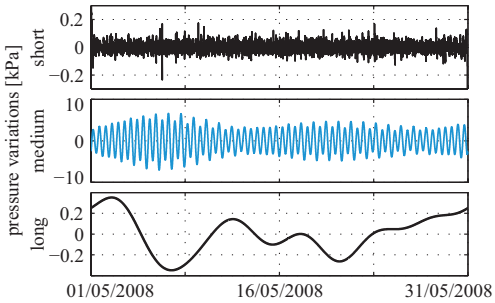
The long term component is best predicted by Sparse Decomposition followed by EMD and EEMD. The results of Harmonic and Wavelet Decomposition are worst due to boundary effects.

### 5.3.2. Decomposition of MAR

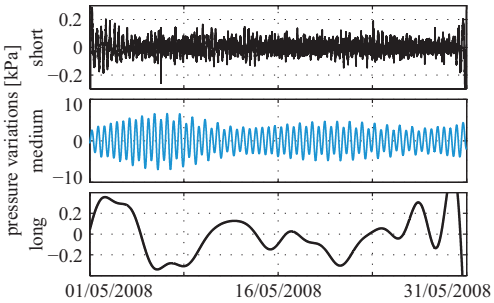
Next, we discuss the results of the second data set MAR, which are given in Figure 5.11. Again, the short and long time components are also shown grouped in Figure 5.12. First of all, we see that this time all algorithms provide a decomposition with short, medium and long time features. Moreover, Sparse Decomposition is again capable to detract the noise from the short time component.

The visible peaks at Harmonic Decomposition and EMD are also extracted by Sparse Decomposition. All these three short time components look quite similar and differ a lot from the short time components obtained by Wavelet Decomposition and EEMD. The short time component of the Wavelet Decomposition shows only one significant peak, has a quite high undefined 'noise level' and huge boundary effects. The worst short time component is given by EEMD. The added white noise does not fade away, hence, no significant short time feature can be obtained. Since this data set is not synthetic, we do not know which results are best. The short time components of Harmonic and Sparse Decomposition as well as EMD look quite similar and these tools are methodically completely different, thus, we guess that the real short time component might be similar to the one provided by this three methods.

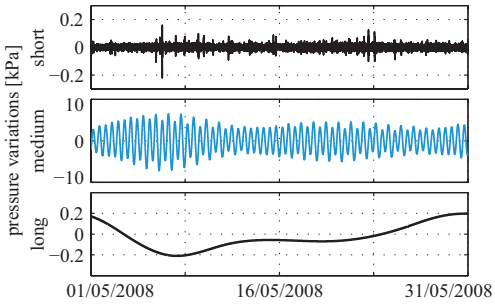
The long time component of all five methods are basically the same even if Har-



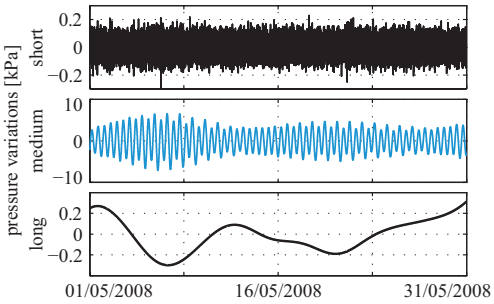

(a) Harmonic Decomposition



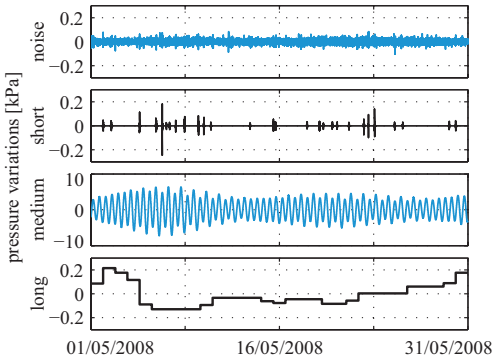

(b) Wavelet Decomposition




(c) EMD



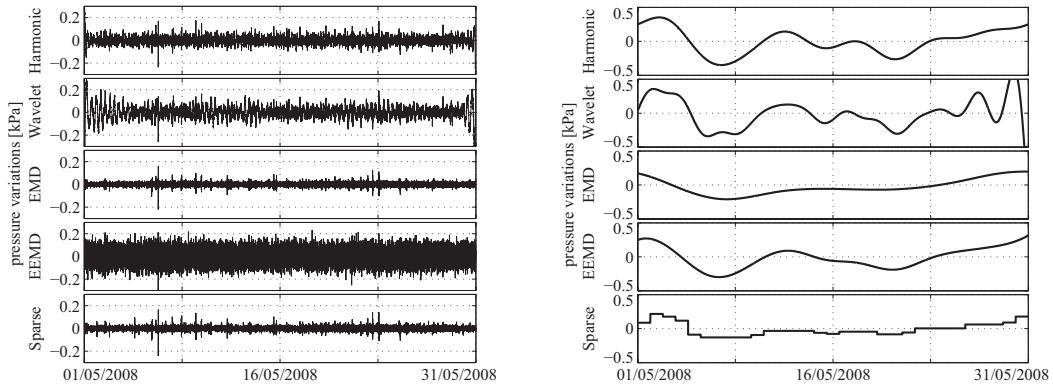

(d) EEMD




(e) Sparse Decomposition

Figure 5.11.: Decomposition of the data set MAR. There are shown from the top to the bottom the short, medium and long time components. Additionally, Sparse Decomposition separates the short time features from the noise.

monic and Wavelet Decomposition as well as EEMD show a lot more waves. All five trends provide a rapid downdrift of about 0.4 $kPa$ during the first week and a slow updrift of the same amount in the following three weeks. The other features obtained in these trends, like the waves in the one obtained by Harmonic and Wavelet Decomposition, are not given significantly in the other methods' trends. Hence, they are probably a byproduct of these methods. As we have seen already in other components, the long time component of the Wavelet Decomposition shows significant boundary effects.




(a) Short time components of MAR

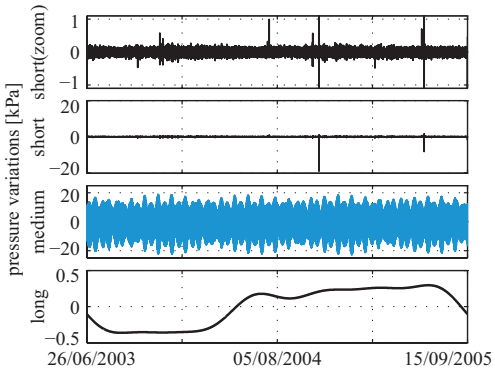
(b) Long time components of MAR

Figure 5.12.: The short (left) and long (right) time components of the data set MAR.

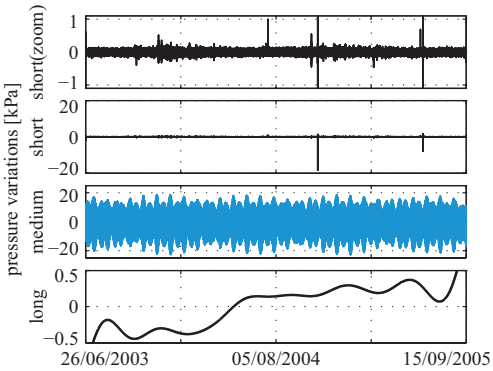
### 5.3.3. Decomposition of CORK1

The results of the first data set at Vancouver Island are presented in Figures 5.13 and 5.14. Again all algorithms provide at least a decomposition in short, medium and long time components and furthermore, Sparse Decomposition is able to separate the noise from the short time component. The short time component of the data set CORK1 contains two events with an amplitude of approximately 20 $kPa$. Thus, we have to have a look at two scales. In addition to the full view of the short time component, we show this component also zoomed in for a more detailed presentation.

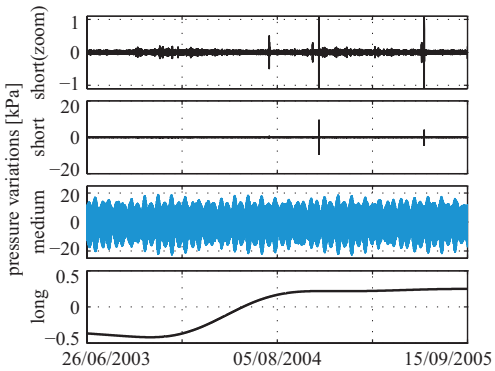
The short time components of Harmonic and Wavelet Decomposition look very similar in any case. The detected peaks are the same and they also do not differ at the basis noise level. All five methods detect the two major short time features at the same location in time but differ in the amplitude of the first event. EMD and EEMD detect these features nearly symmetrically to the time axis. This might be of methodical reason since the resulting IMFs are often similar to harmonic waves. As at the data set MAR, the short time component of EEMD is not very useful since the amplitude of the remaining added noise is of the same scale as the



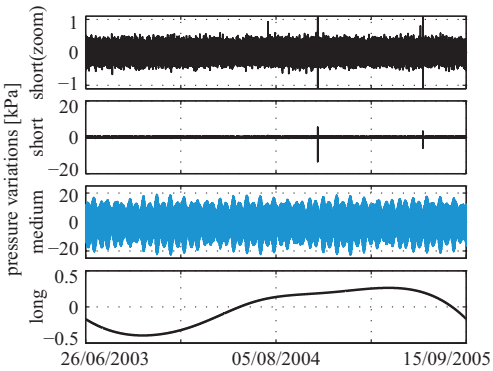

(a) Harmonic Decomposition



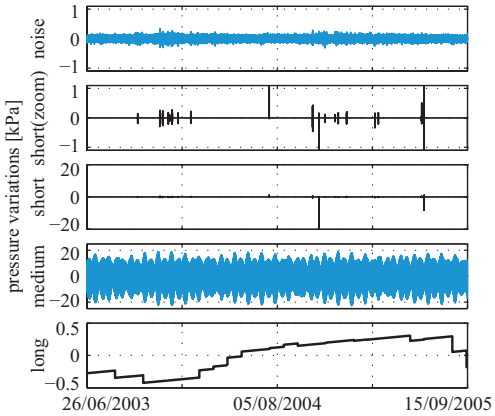

(b) Wavelet Decomposition




(c) EMD

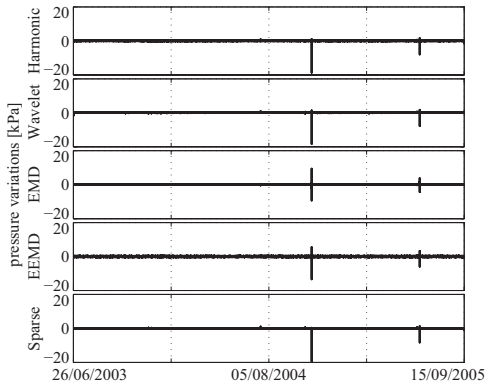



(d) EEMD

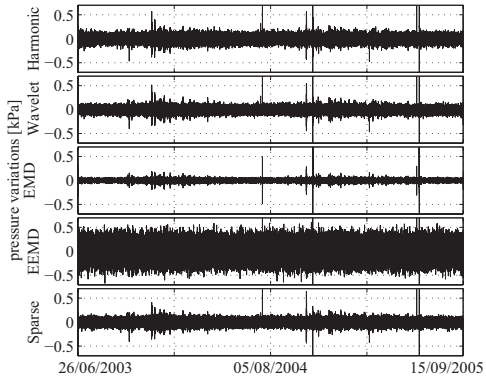



(e) Sparse Decomposition

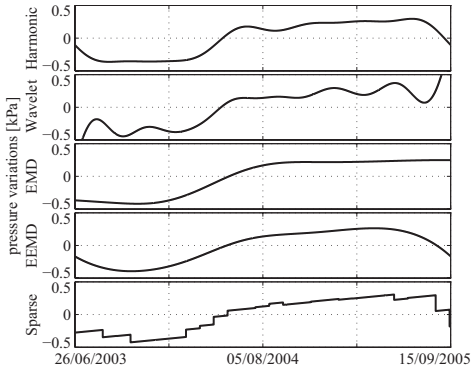
Figure 5.13.: Decomposition of the data set CORK1. There are shown from the top to the bottom the short, medium and long time components.




(a) Short time components of CORK1




(b) Short time (zoom) components of CORK1

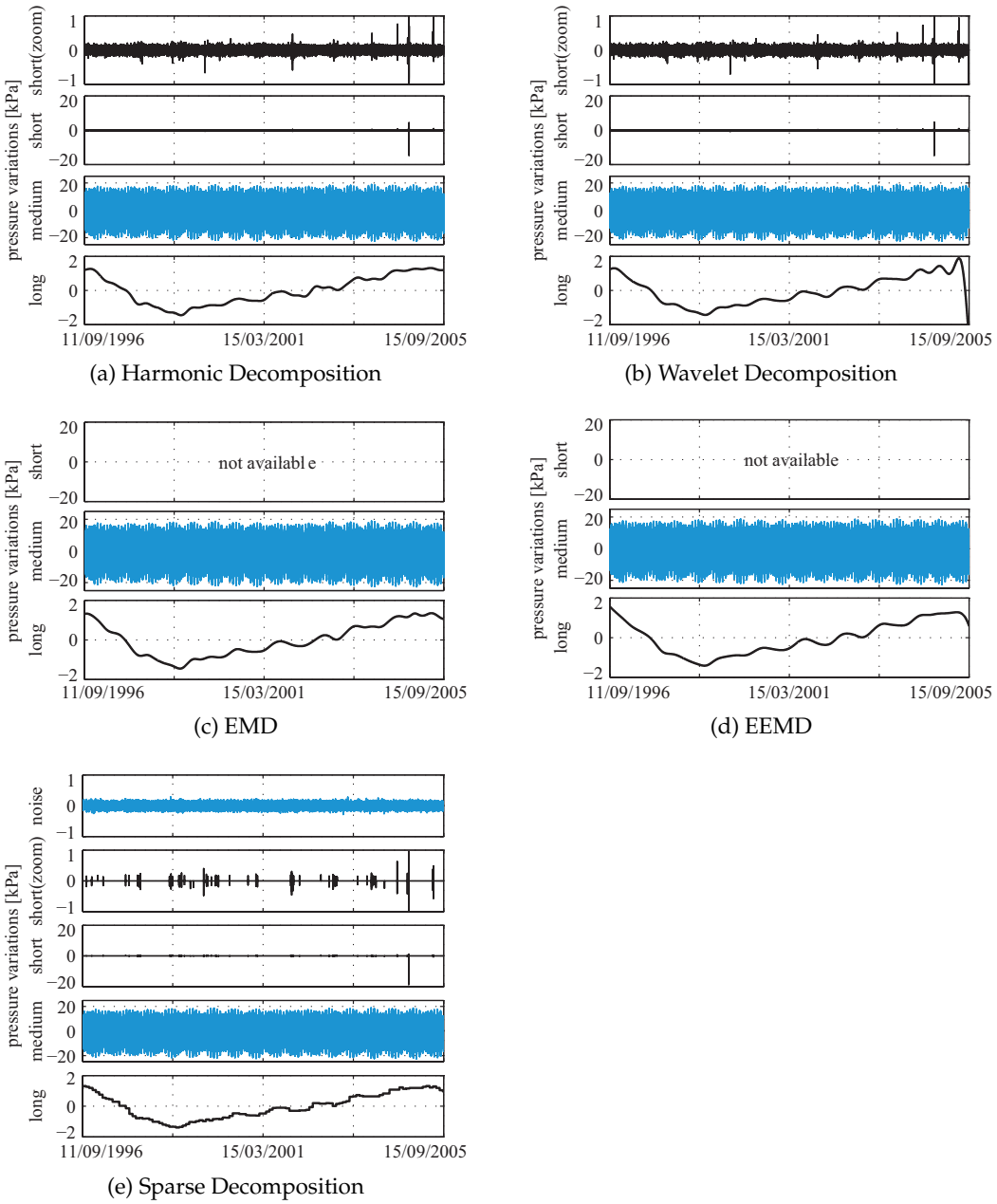



(c) Long time components of CORK1

Figure 5.14.: On the left hand side are the short time components of CORK1 decomposed by the five methods. On the top showed in the usual way and on the bottom zoomed in. On the right hand side are the long time components of CORK1.

events we are looking for.

On the contrary to the dissimilarity of the short time components, the long time components of the five methods look very similar. They all detect an almost constant or a slightly decreasing pressure variation from June 2003 to spring 2004 followed by an updrift of around 0.7 $kPa$ in spring 2004. Afterwards, there is again a period of steady pressure variation, which lasts around one year. At the end of the measuring interval, Harmonic and Sparse Decomposition as well as EEMD detect a decrease of around 0.4 $kPa$. On the opposite, EMDs long time component does not show a decrease and the Wavelet Decomposition is highly influenced by boundary effects.




(a) Harmonic Decomposition

(b) Wavelet Decomposition

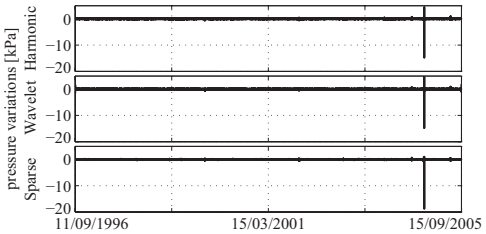
(c) EMD

(d) EEMD

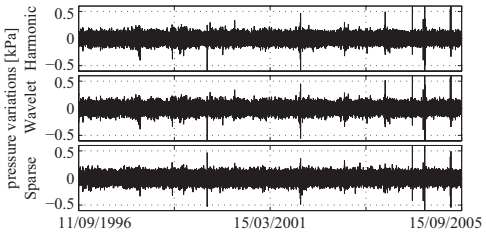
(e) Sparse Decomposition

Figure 5.15.: Decomposition of the data set CORK2. There are shown from the top to the bottom the short, medium and long time components. Since the short time components has features of different amplitudes these are also showed zoomed in. EMD and EEMD failed to detect the short period features.

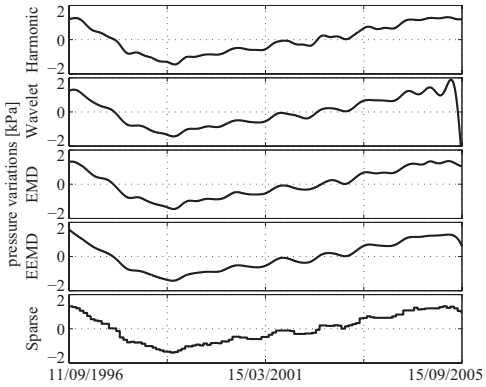
### 5.3.4. Decomposition of CORK2




(a) Short time components of CORK2




(b) Short time (zoom) components of CORK2




(c) Long time components of CORK2

Figure 5.16.: The short time components of CORK2 are on the left hand side. On the top in the usual way and on the bottom zoomed in. On the right hand side are the long time components of CORK2.

At last, we present the result of the second data set at Vancouver Island. This is with a duration of more than 9 years the longest data set. The results are shown in Figures 5.15 and 5.16. Since the sampling interval is only 60 *min*, we can not expect short time features with a shorter duration than 60 *min* to appear. Hence, the short time components have to be interpreted very carefully. As we have seen already at the data set SYN, EMD and EEMD are again not capable to provide a short time components for this data set. The other three methods separate a short time component and in addition, Sparse Decomposition separates the noise.

The long time components look again very similar. All five methods detect a decrease of about 3.5 *kPa* in the first two years and an increase of the same amount in the following seven years. Additionally, all long time components show waves of period around one year especially at the ramp part of the trend, which might be due to the long time tidal constituents. As already seen very often at the other data sets the long time component of CORK2 created by Wavelet Decomposition is affected by significant boundary effects.

### 5.3.5. Discussion

We start our discussion by having a look at the computing time of the methods which are shown in Table 5.5.

Table 5.5.: Summary of the speed of the methods measured in CPU time; *: efficiently implemented by Matlab®; **: without creating the operator

| | | Harmonic* | Wavelet* | EMD | EEMD | Sparse** |
|---|---|---|---|---|---|---|
| CPU time | SYN | 0.001 s | 1.09 s | 0.21 s | 36.58 s | 1.40 s |
| | MAR | 0.027 s | 1.89 s | 30.30 s | 15.03 min | 10.26 min |
| | CORK1 | 0.107 s | 2.70 s | 12.00 min | 9.86 h | 19.20 h |
| | CORK2 | 0.103 s | 2.48 s | 19.77 s | 54.81 min | 8.33 days |

The results of Harmonic and Wavelet Decomposition as well as EMD and EEMD were obtained using an AMD Athlon(tm) 64 X2 Dual Core Processor 3800+ with 2 GHz per core and in total 1.95 GB RAM. Since the computing effort for Sparse Decomposition is considerably longer we had to use an external computer with two Quad-Core AMD Opteron(tm) Processor 2376, 2.3 GHz per core, and in total 15.7 GB RAM. The current version of RFSS can not be scheduled in parallel, thus, we could only use one core.

Clearly, the computing time of Harmonic and Wavelet Decomposition are the best, i.e. they need the smallest amount of time, but to be fair they are based on the Matlab® routines `fft` and `wavedec`, which are highly efficient algorithms and implemented very fast. They are almost unaffected by the number of measurements of the data set and do not need more than 3 seconds for any of our data sets. The other algorithms are a lot more complex, thus, computing time is a limiting factor if using these methods. EMD, which needs some seconds to several minutes depending on the length of the data set, is also quite applicable to long data sets. The algorithms EEMD and RFSS, which is the major step of Sparse Decomposition, are a lot slower if applied to large data sets. The running time of EEMD goes up to several hours and the one of RFSS up to several days.

We complete the discussion of the results by summarising the previous section in Table 5.6 based on the following criteria. First, how successful was the algorithm in decomposing the signal into a long and short period components, especially, did the algorithm find a short period event in the presence of background noise? Second, does the algorithm create large effects at the boundaries of the time window? Lately, how much CPU time is needed for the decomposition? In most cases we do not know what the real decomposition is and which features are really due to boundary effects. Furthermore, the algorithms are written in different programming languages and are run on different computers. Thus, we decided to judge only in three categories, namely good, fair and bad. Of course due to this coarse judgement there might be the case that two methods with the same mark and one performed clearly better. But we prefer to make this fault and avoid to overstate the results. The category 'computing efficiency' was evaluated due to

the running times of the algorithms shown in Table 5.5.

Table 5.6.: Overview of the results. ✔ indicates good, ● fair, and ✗ bad results.

| | | | Harmonic | Wavelet | EMD | EEMD | Sparse |
|---|---|---|---|---|---|---|---|
| Decomposition | short period | SYN | ✔ | ✔ | ✗ | ✗ | ✔ |
| | | MAR | ● | ● | ✔ | ✗ | ✔ |
| | | CORK1 | ✔ | ✔ | ● | ● | ✔ |
| | | CORK2 | ✔ | ✔ | ✗ | ✗ | ✔ |
| | long period | SYN | ● | ● | ✔ | ● | ✔ |
| | | MAR | ✔ | ● | ✔ | ✔ | ✔ |
| | | CORK1 | ✔ | ● | ✔ | ✔ | ✔ |
| | | CORK2 | ● | ● | ✔ | ✔ | ✔ |
| Boundary | effects | SYN | ✗ | ✗ | ✔ | ● | ✔ |
| | | MAR | ● | ✗ | ✔ | ✔ | ✔ |
| | | CORK1 | ● | ✗ | ✔ | ✔ | ✔ |
| | | CORK2 | ● | ✗ | ✔ | ✔ | ✔ |
| Computing efficiency | | | ✔ | ✔ | ● | ✗ | ✗ |

First of all, everyone can notice by looking at Table 5.6 that no method is perfect in every aspect.

Harmonic and Wavelet Decomposition perform very well in short period features and quite well in trend detection but their results show massive boundary effects. Also mentionably is that Harmonic and Wavelet Decomposition are really fast even if the number of measurements goes up to several hundred thousands.

EMD and EEMD have problems by decomposing the short period features when the sampling interval is 60 min. The added white noise by EEMD, which should improve the EMD, does not fade away in one houndred runs and is clearly a disadvantage when trying to detect components with a small amplitude. In contrast to that, both EMD and EEMD perform very well in detecting long time components. Another drawback is the additional computational effort of the EEMD which is, by comparing the results of EMD and EEMD, not worthwile.

Sparse Decomposition behaved very well in detecting short and long period features and furthermore, the noise was separated at any data set. Another advantage of Sparse Decomposition is that we did not obtain any boundary effects at the computed decompositions. The only disadvantage is the computing efficiency, which clearly needs to be improved for practical purposes.

# 6 Conclusions

In this thesis we applied the notion of sparsity to decompose sea floor pressure data sets into components of similar scale in time. This is achieved by $\ell^1$ minimisation with an overcomplete dictionary. Since the corresponding dictionary operator is overcomplete and some submatrices in our algorithm are not invertible we needed a stabilised $\ell^1$ penalised Tikhonov functional. It turned out that the elastic net, which combines a $\ell^1$ and a $\ell^2$ penalty term, fulfils the needed requirements. An efficient way to solve the elastic net is given by the RFSS. This algorithm iteratively solves the optimal conditions of the elastic net. These are given by using the subdifferential calculus.

Further investigations have shown that the elastic net has a unique minimiser as long as the stability parameter is non-zero. In addition, we have shown that the minimiser depends continuously on the parameters. This main result is exploited to derive further smoothness results for similar parameter-dependent mappings.

The application of Sparse Decomposition to the four sea floor pressure data sets have shown that this is a appropriate way to decompose this data sets into physical meaningful components. We also tested four other decomposition methods, namely the classical Harmonic and Wavelet Decomposition as well as the novel Empirical Mode Decomposition and its enhancement the Ensemble Empirical Mode Decomposition. These methods also showed some useful properties. The classical methods are very fast and their decompositions are acceptable. In particular, the Harmonic decomposition has shown good quality, since it is not as susceptible to boundary effects as the Wavelet Decomposition. The novel method EMD is also acceptable. Its computing effort is moderate and especially its long time components were reasonable even if it does not detect sharp features as steps appropriately. Only the ability to decompose the short time component was not decent. The enhancement of the EMD, EEMD, performed never better than the EMD. It never showed that the additional computing effort is worthwhile. Also the added white noise, which does not fade away, degrades the results relevantly.

Overall, Sparse Decomposition performs best in decomposing sea floor pressure data sets at least for short data sets, i.e. less than about 100,000 measurements. For larger data sets we have to use the other techniques or to enhance this method for longer data sets. This can be done either by speeding up the RFSS or by finding better dictionaries, since the sparsity of the data set in the dictionary is crucial for the speed.

For further studies on this field the following issues are of interest. On the one hand, there are some issues concerning the parameters of the elastic net. If the data is in the image of the dictionary operator, which is the case for overcomplete dictionaries, how do we have to choose the sparsity parameter to get a decomposition so that the residual is less than any predefined tolerance threshold? What is a proper choice of the stability parameter?

On the other hand, there are also questions left concerning the speed of the RFSS. How fast is the RFSS in average? Is it possible to proof a rate of convergence if we make assumptions on the dictionary operator or the data? How can we speed up the RFSS? Is there any modification of the RFSS so that we can use parallel computing for speed up?

# List of Notations

| | |
|---|---|
| $A_c, \mathring{A}, \overline{A}$ | complement, interior and closure of a set $A$ |
| $\mathfrak{P}(A)$ | powerset of a set $A$ |
| $\mathrm{Im}(f)$ | image of $f : X \to Y$, $\mathrm{Im}(f) := \{y \in Y : \exists\, x \in X \text{ s.t. } f(x) = y\}$ |
| $\mathbb{R}_\infty$ | the extended real numbers, $\mathbb{R}_\infty := \mathbb{R} \cup \{\infty\}$ |
| $\mathcal{L}(X, Y)$ | linear and continuous mappings from $X$ to $Y$ |
| $X', X''$ | the dual space and second dual space of $X$ |
| $\langle x', \cdot \rangle$ | duality pairing, $\langle x', \cdot \rangle := x'(\cdot), x' \in X'$ |
| $K'$ | adjoint operator to $K \in \mathcal{L}(X, Y), K' : Y' \to X'$ |
| $\partial_\alpha f, \partial_n f$ | partial derivatives, $\partial_\alpha f(x) := \frac{\partial f}{\partial \alpha}(x)$, $\partial_n f(x) := \frac{\partial f}{\partial x_n}(x)$ |
| $\mathcal{G}_f(x)$ | Gâteaux derivative of $f$ at $x$ |
| $\mathcal{F}_f(x)$ | Fréchet derivative of $f$ at $x$ |
| $\partial f$ | subdifferential of $f : X \to \mathbb{R}_\infty$, $\partial f : X \rightrightarrows X'$ |
| $\rightrightarrows$ | set valued mapping |

**elastic net**

| | |
|---|---|
| $\Phi_{\alpha,\beta}$ | elastic net, $\Phi_{\alpha,\beta}(x) := \frac{1}{2}\|\mathcal{D}x - y\|^2_{\mathcal{H}} + \alpha\|x\|_{\ell^1} + \frac{1}{2}\beta\|x\|^2_{\ell^2}$ |
| $\ell^p$ | space of sequences, which satisfy $\sum_{n=1}^{\infty} \|x_n\|^p < \infty$, $p < \infty$ |
| $\mathbb{R}^+$ | $\mathbb{R}^+ := \{t \in \mathbb{R} : t > 0\}$ |
| $\mathbb{R}^+_0$ | $\mathbb{R}^+_0 := \mathbb{R}^+ \cup \{0\}$ |
| $\mathbb{P}$ | set of all pairs of parameters, $\mathbb{P} := \mathbb{R}^+_0 \times \mathbb{R}^+$ |
| $\alpha$ | sparsity parameter |
| $\beta$ | stability parameter |
| $y$ | data |
| $\mathcal{D}$ | dictionary operator |

**RFSS**

| | |
|---|---|
| $M, N$ | finite dimensions, $\mathcal{D} : \mathbb{R}^M \to \mathbb{R}^N$ |
| $\mathcal{N}$ | set of all possible indices, $\mathcal{N} := \{1, \ldots, N\}$ |
| $\Gamma$ | active set, the set of all indices $n$ so that $x_n \neq 0$, $\Gamma \subset \mathcal{N}$ |
| $x_\Gamma$ | restriction of $x = (x_n)_{n\in\mathcal{N}}$ to the active set, $x_\Gamma := (x_n)_{n\in\Gamma}$ |
| $\lvert x\rvert_n$ | modulus of the $n$th component, $\lvert x\rvert_n := \lvert x_n\rvert$ |
| $d_n$ | a pattern; a column of the dictionary operator $\mathcal{D}$ |
| $\mathcal{D}_\Gamma$ | restriction of the dictionary operator $\mathcal{D} = (d_n)_{n\in\mathcal{N}}$ to the active set, $\mathcal{D}_\Gamma := (d_n)_{n\in\Gamma}$ |
| $\Xi_{\theta,\Gamma}$ | auxiliary functional, $\Xi_{\theta,\Gamma}(x) := \frac{1}{2}\lVert\mathcal{D}_\Gamma x_\Gamma - y^\delta\rVert_2^2 + \alpha\langle\theta_\Gamma, x_\Gamma\rangle + \frac{1}{2}\beta\lVert x_\Gamma\rVert_2^2$ |
| $\Xi^k$ | abbreviation of the auxiliary functional, $\Xi^k := \Xi_{\theta^k,\Gamma^k}$ |
| $n^*$ | added index in 'next pattern' |
| $m^*$ | removed index in 'line search' |
| $(\Gamma, x, \theta)^k$ | $k$th triple, $(\Gamma, x, \theta)^k := (\Gamma^k, x^k, \theta^k)$ |
| $\nabla_\Gamma f$ | gradient with respect to the active set, $\nabla_\Gamma f := (\partial_n f)_{n\in\Gamma}$ |
| $R(x)$ | residual, $R := \mathcal{D}x - y$ |

# Bibliography


Alt, H.W., 2006. Lineare Funktionalanalysis: Eine anwendungsorientierte Einführung (German Edition). Springer. 5. edition.

Bredies, K., Lorenz, D.A., 2009. Regularization with non-convex separable constraints. Inverse Problems 25.

Bredies, K., Lorenz, D.A., 2011. Mathematische Bildverarbeitung: Einführung in Grundlagen und moderne Theorie (German Edition). Vieweg + Teubner.

Chen, S.S., Donoho, D.L., Saunders, M.A., 1999. Atomic decomposition by basis pursuit. SIAM journal on scientific computing 20, 33–61.

Daubechies, I., Defrise, M., De Mol, C., 2004. An iterative thresholding algorithm for linear inverse problems with a sparsity constraint. Communications on Pure and Applied Mathematics 57, 1413–1457.

Elstrodt, J., 2004. Maß- und Integrationstheorie (German Edition). Springer. 4. edition.

Gröchenig, K., 2001. Foundations of Time-Frequency Analysis. Applied and Numerical Harmonic Analysis, Birkhäuser Boston Inc., Boston, MA.

Huang, N.E., Shen, Z., Long, S.R., Wu, M., Shih, H.H., Zheng, Q., Yen, N.C., Tung, C.C., Liu, H.H., 1998. The Empirical Mode Decomposition and the Hilbert Spectrum for Nonlinear and Non-Stationary Time Series Analysis. Proceedings of the Royal Society of London. Series A: Mathematical, Physical and Engineering Sciences 454, 903–995.

Jin, B., Lorenz, D.A., Schiffler, S., 2009. Elastic-net regularization: Error estimates and active set methods. Inverse Problems 25.

Lee, H., Battle, A., Raina, R., Ng, A.Y., 2007. Efficient sparse coding algorithms. Advances in neural information processing systems 19.

Louis, A.K., Maaß, P., Rieder, A., 1997. Wavelets: Theory and Applications. Wiley: New York.

Pawlowicz, R., Beardsley, B., Lentz, S., 2002. Classical Tidal Harmonic Analysis Including Error Estimates in MATLAB using T_TIDE. Computers and Geosciences 28, 929–937.

Rilling, G., 2007. EMD-algorithm, `emd.m`. `http://perso.ens-lyon.fr/patrick.flandrin/emd.html` .

Rockafellar, R.T., 1972. Convex Analysis. Princeton University Press.

Rockafellar, R.T., Wets, R.J.B., 1991. Variational Analysis. Springer.

Rudin, W., 1991. Functional Analysis. McGraw-Hill Inc.

Schiffler, S., 2009. RFSS-algorithm, `rfss.m`. `http://sites.google.com/site/igorcarron2/cs/` .

Schiffler, S., 2010. The elastic net: Stability for sparsity methods. Ph.D. thesis. University of Bremen.

Stark, H.G., 2005. Wavelets and Signal Processing: An Application-Based Introduction. Springer. 1 edition.

Stewart, H.R., 1997. Introduction to Physical Oceanography. Texas A. & M. University.

Werner, D., 2007. Funktionalanalysis (German Edition). Springer, Berlin [u.a.]. 6. edition.

Wu, Z., downloaded 04/2011. EEMD-algorithm, `eemd.m`. `http://rcada.ncu.edu.tw/eemd.m` .

Wu, Z., Huang, N.E., 2009. Ensemble Empirical Mode Decomposition: A Noise-Assisted Data Analysis Method. Advances in Adaptive Data Analysis (AADA) 1, 1–41.