National Technical University of Athens
School of Applied Mathematical and Physical Sciences

Sparse recovery and matrix completion

Mathematical aspects and algorithms

Ioannis C. Tsaknakis

Supervisor : Michail Loulakis
Associate Professor, NTUA

A thesis presented for the degree of
Master of Science in Applied Mathematical Sciences

Athens, Greece
July 2017

National Technical University of Athens
School of Applied Mathematical and Physical Sciences

Sparse recovery and matrix completion

Mathematical aspects and algorithms

Ioannis C. Tsaknakis

BSc., Electrical and Computer Engineering
National Technical University of Athens (2015)

Supervisor : Michail Loulakis
Associate Professor, NTUA

(Signature)	(Signature)	(Signature)
.....	.....	.....
Michail Loulakis	Dimitris Fouskakis	Filia Vonta
Associate Professor, NTUA	Associate Professor, NTUA	Associate Professor, NTUA

A thesis presented for the degree of
Master of Science in Applied Mathematical Sciences

Athens, Greece
7-7-2017

Acknowledgements

First of all, i would like to thank associate professor Michalis Loulakis for his continuous help during the months i worked on this thesis, as we had meetings twice a month to discuss my progress. Also, i am grateful to him for the freedom he gave me with respect to the way i work and the subject of this thesis. He was willing to help me although the main subject of this thesis was outside of his main research interests.

Furthermore, this thesis would not be complete without the contribution of the team at the Institute for Astronomy, Astrophysics, Space Applications and Remote Sensing (IAASARS) of the National Observatory of Athens (NOA). Specifically, i would like to thank researchers Konstantinos Koutroumbas and Athanasios Rontogiannis, as well as doctoral student Paris Giampouras. I am grateful to them for the useful discussions we had and the time they spend with me at the IAASARS site in Penteli helping me develop chapter 6.

Abstract

The "data deluge" we are facing necessitates us to understand the structure of the data we acquire and process. For that reason we introduce the notion of *low-dimensional models*, i.e subsets of the signal space (the space in which the data/signals reside) with specific properties (structure). We restrict our attention on two signal models : sparse vectors and low-rank matrices. In sparse vector models we focus on the *sparse vector recovery problem*. This problem is about reconstructing/recovering a sparse signal after projecting it in a lower dimensional space. The important result we present and prove is the ability, under certain conditions, of certain classes of random matrices, such as Gaussian and subgaussian random matrices, to project sparse vectors to lower dimensional spaces in a way such that we can obtain successfully the original signal vectors (reconstruction) using suitable algorithms. In low-rank matrices we focus on a special recovery problem called *matrix completion*. In that problem we are given an incomplete matrix and the prior knowledge that this matrix is low-rank and we want to infer the values of the missing entries. In the last part of this thesis we introduce several algorithms for the matrix completion problem and assess them in synthetic and real data.

Keywords

Big Data, low-dimensional models, sparse vectors, low-rank matrices, sparse recovery, matrix completion, l_1 minimization, nuclear norm minimization, random projections, Gordon's lemma, concentration of measure, singular value thresholding

Preface

The massive amounts of data that are produced nowadays are creating huge challenges for the modern computational systems. The great impact of information technologies in all aspects of our lives make imperative the discovery of novel approaches for solving these problems. The recent years understanding the structure of data we are dealing with and exploiting that knowledge in order to tackle the emerging challenges has proven to be a productive and effective approach.

The main concept behind this thesis is the importance of exploiting the structure of data for performing information processing tasks efficiently and successfully. We restrict our study on two specific structures, which we will later call low-dimensional signal models : sparse signals and low-rank matrices. Specifically we will focus our attention on two important problems of these models : sparse recovery and matrix completion.

In my opinion one of the most fascinating aspects of machine learning is that it manages to blend in a fruitful and interesting way theoretical (purely mathematical) and applied (practical) topics. The main topic of this thesis (sparse recovery and matrix completion) lies in the intersection of machine learning, signal processing and applied mathematics and is a characteristic example of such an interesting blend. For that reason this thesis deals both with mathematical and applied issues of the above problems. This is reflected in the structure of this thesis, since part 2 surveys some advanced mathematical topics of sparse recovery, while in part 3 several matrix completion algorithms are described and evaluated on synthetic and real data.

Specifically, this thesis consists of three parts:

- The first part is essentially a survey of the the two signal models of interest, the sparse vectors model and the low-rank matrices model.
  - The first chapter contains a general description of *low-dimensional signal models*, the reason we study them and the notion of stable embedding. This chapter highlights the importance of understanding the underlying structure of data.
  - The second chapter is devoted to *sparse vectors*. We start with the necessary definitions and the general description of the *sparse vector recovery problem* and move to topics such as the optimization tasks of sparse recovery, restricted isometry property, algorithms for sparse recovery and applications.
  - The third chapter is about *low-rank matrices*. Besides providing the necessary background, we focus on the problem of matrix completion.

- The second part contains some more advanced mathematical topics of the sparse vector recovery problem. Essentially, the theme of this chapter is the ability of certain classes of random matrices to project sparse vectors to lower dimensional spaces in a way such that we can obtain successfully the original signal vectors (reconstruction) using suitable algorithms.
  - The fourth chapter contains the necessary probabilistic background. It starts with basic notions and definitions and moves to more advanced topics, such as concentration of measure, suprema of stochastic processes, Gaussian width, etc.
  - The fifth chapter contains two main results along with the respective proofs. Both results refer to the ability, under certain conditions, of certain classes of random matrices to project sparse vectors in a way such that successful recovery is guaranteed using suitable algorithms. The first result deals with Gaussian random matrices, while in the second result we refer to subgaussian matrices.
- The third part studies and evaluates five different matrix completion algorithms.
  - The sixth chapter introduces five different optimization tasks for the matrix completion problem and the respective algorithms that solve them. Then it proceeds by evaluating the algorithms on synthetic and real data (MovieLens dataset).

Contents

Part I

Sparse vectors and low-rank matrices

Chapter 1

Introduction

1.1 Introduction

We live in the *Big data era*. Huge amounts of data are produced every day and that amount is continuously increasing. Based on previous estimates [Ibm] that amount is larger than 2.5 exabytes per day and is increasing exponentially. These data can take a variety of forms, such as image, video, audio, text and can be found practically anywhere (social networks, radio telescopes, DVDs, medical records, etc.). That tremendous amount of data is putting a strain on our systems, a situation that is going to become more demanding the forthcoming years. In order to get a good grasp of the magnitude of data we are referring to, figure 1.1 presents the expected amount of data that are going to be used in 2025 for acquisition and storage in three different data domains (astronomy, genomics and YouTube), as given in [Ste+15].

The data we are referring to, the "Big Data", have several characteristics. Among those characteristics is the high dimension of the space they reside and the their huge volume. Furthermore, those data are usually grossly corrupted due to noise, failure of the sensing devices, or intentional tampering. The situation becomes more complicated since some elements of those data are usually missing; their volume and dimensionality makes it nearly impossible to collect them all. It is apparent that these traits of data makes the situation even more challenging.

The enormous flow of data we are dealing with, known as *data deluge*, and their complexity, is creating several challenges concerning the acquisition, storage, processing, learning and transmission of data. The importance of data in nearly every aspect of our lives and the fact that they play a vital role in many different fields such as science, finance and medicine is creating a demand for pushing further the capabilities of our technologies. It is essential to develop

Phase \Domain	Astronomy	Genomics	YouTube
Acquisition	25 ZB/year	1 zetta-bases/year	500-900 million hours/year
Storage	1 EB/year	2-40 EB/year	1-2 EB/year

Figure 1.1: Expected amount of data that are going to be used in 2025 for acquisition and storage in three different data domains. [Ste+15]

novel computational methods, algorithms, mathematical tools and technologies, in order to tackle efficiently all the problems that stem from the data deluge. One such approach is to try to understand the underlying structure of data and exploit that to perform all the necessary tasks successfully and efficiently. This direction serves as the motivating concept behind this thesis.

1.2 Low-dimensional signal models

The fundamental notion we are going to deal with is that of a *signal*. A signal is a function that contains information about the attributes or the behavior of a system or some phenomenon [Pri90] and can refer to practically any time or space varying quantity, such as music, audio, video, images, financial data, scientific data, text, etc. We can also refer to signals as data. In the context of signal processing we use the term signal, while in the machine learning community the term *data* is preferred. Nevertheless, these terms both refer to some piece of information that we want to process (e.g compress, code, learn, denoise, etc.). In this thesis the main subject lies in the intersection of machine learning and signal processing, so we are going to refer to information patterns, either as signals or data, depending on the context without causing confusion.

The signals we study reside in a mathematical space, known as *signal space*. In other words, the signal space is the set of all possible values the signal we study can take. Sometimes, for reasons that will become apparent a bit later, it is necessary to restrict our attention to a subset of the signal space. A *signal model* is a subset of the (universal) signal space that possesses certain properties, usually expressed in the form of mathematical statements. A (low-dimensional) signal model is characterized by a number of degrees of freedom that is smaller than the respective number for the signal space.

There are several types of low-dimensional signal models. The most notable of them are the following. [BCW10]

1. Sparse vectors
2. Compressible vectors
3. Low-rank matrices
4. Manifolds
5. Point clouds

In a nutshell, a *sparse vector* is a vector that has at most k non-zero elements, where k is a relatively small number compared to the dimensionality of the ambient space. This is the most important and well-studied model and we are going to present it in detail in chapter 2.

A *compressible vector* is a generalization of the sparse signal model, where we have k significant elements (i.e elements with large values), while the rest of the elements are very small. We are going to present the basics of this model in chapter 2.

A *low-rank matrix* is a matrix whose rank is small compared to it's dimension. It is an interesting model with many applications. It will comprise the subject of chapter 3.

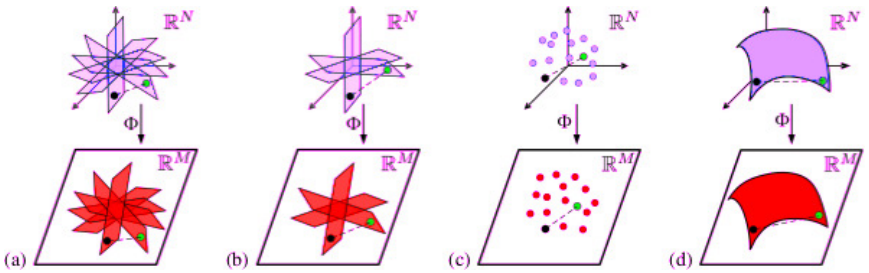

Figure 1.2: A collection of low-dimensional signal models: (a) k-sparse signals, (b) structured k-sparse signals, (c) a point cloud, (d) a smooth k-dimensional manifold. Image taken from [BCW10].

A *manifold*, from a mathematical point of view, is a topological space that resembles locally the Euclidean space. Essentially, we can consider a manifold as a generalization of a surface.

A *point cloud* is collection of a finite number of points in $\mathbb{R}^n$, i.e vectors $\boldsymbol{x} \in \mathbb{R}^n$ that represent signals.

Figure 1.2 illustrates four characteristic low-dimensional signal models, namely sparse vectors, structured sparse vectors [1], point clouds and manifolds. [BCW10]

1.2.1 Inverse problems

Low-dimensional models led to significant advances in the solution of linear inverse problems. *Inverse problems* [The15] are problems, where a set of input-output observations were given (training data in machine learning) and the objective is to estimate or predict the parameters of the model that produces that set of observations. Roughly, in inverse problems the task is to infer the cause from the effect. In general, inverse problems are *ill-posed*. That means that at least one of the following three conditions is violated : existence, uniqueness and stability of solution. Finding the solution is in many cases a difficult, if not impossible, task. The reason is that the model we use to describe the data is too complex. As a result, it requires a number of parameters that is too large, usually larger than the number of data points. Therefore, we cannot estimate the parameters of the model from our observations.

The study of low-dimensional models led to a breakthrough in inverse problems. Exploiting the fact that the model of our data possesses a special structure, i.e the number of parameters that describe it is smaller than the number of parameters characterizing the ambient space, we only need a small number of data points in order estimate the parameters of the underlying model, which is the final objective. It is typical in inverse problems to use as a prior assumption that our model possesses a special structure. Note that this assumption is based on the properties of the respective physical system, which the model describes.

Inverse problems are ubiquitous in science and engineering. The majority of problems in signal processing and machine learning fall under the scope of inverse problems. Typical examples are problems as regression, classification, signal denoising and many more.

[1] Structured sparse vectors are a special case of sparse vectors, where the positions of the non-zero elements are not completely arbitrary, but follow a specific pattern.

1.3 Stable embeddings

One of the most important concepts in machine learning is that of *dimensionality reduction*. It refers to algorithms that project data to a lower (than the original) dimensional space, essentially reducing the number of variables that describe the data. The concept of projection (dimensionality reduction) relies on the fact that the some signals can be (approximately) described using less parameters than the number of parameters used to describe an arbitrary signal of the ambient signal space. We should note that formally, a *projection* is a linear function [2] $\Phi : \mathbb{R}^n \to \mathbb{R}^m$, $m < n$. The main motivations behind employing dimensionality reduction is the speedup of the data processing and the reduced requirements in storage space and memory.

Essentially, the objective is to design a suitable projection such that the low-dimensional signal retains a significant amount (in the ideal situation all) of the information content of the original signal. Generally, that is impossible, as for the projection matrix it holds that $null(\Phi) \neq \{0\}$ (when $m < n$, which in our case holds). As a result, several signals share the same low-dimensional projection, i.e $\Phi \cdot \boldsymbol{x} = \Phi(\boldsymbol{x} + \boldsymbol{z}), \forall \boldsymbol{z} \in null(\Phi)$. Consequently we need to consider subsets of the signal space, i.e signal models, hoping that we will be able to construct projections that preserve information in these models. In order to construct these projections it is important to be able to describe the ability of projections to preserve the information content of signals. One such notion is that of a *stable embedding*.

Definition 1.3.1 (Stable embedding)**.** *[BCW10] Let $U = \mathbb{R}^n$ be the signal space, $\boldsymbol{x} \in U$ an arbitrary signal and $S \subseteq U$ a signal model. A stable embedding on S is a projection $\Phi : \mathbb{R}^n \to \mathbb{R}^m$, $m < n$ such that the following two conditions hold*

1. $\Phi \cdot \boldsymbol{x}_1 \neq \Phi \cdot \boldsymbol{x}_2, \forall \boldsymbol{x}_1, \boldsymbol{x}_2 \in S$
2. *It approximately preserves the distances between all the points of S.*

Essentially, a stable embedding is a projection that approximately preserves the distances between all the points of a signal model [BCW10]. This property of stable embeddings results to robustness to noise. We want to construct stable embeddings for certain classes of signal models, since in the general case this is impossible. An important direction towards that goal is to use random matrices, which will constitute a major part of this thesis.

The main questions that emerge in the development of stable embeddings for signal models are the following. [BCW10]

- What is the smallest dimension m of the lower dimensional space for which we can attain a stable embedding?
- How to construct the proper projection matrix, that provides the stable embedding?

We are going to answer these questions for the sparse signals model in the next chapters. Specifically, we are going to provide bounds for the values of m, with

[2] We restrict our attention to linear projections, although non-linear projections are used in data sciences. The reason is simplicity, since in this part our aim is to highlight the conceptual aspects of the topic

respect to the other parameters of the problem (e.g level of sparsity). Also, we are going to establish that several classes of random matrices provide, under certain conditions, a stable embedding for sparse signals.

Consequently, we can say that one of the main concepts underlying this thesis is the power of randomness to provide information preserving embeddings for specific classes of signal models. This also highlights another important aspect, that of discovering and exploiting the underlying structure of data, i.e finding the low-dimensional model that describes adequately the data. In a nutshell, *random projections* and *low-dimensional structure* are two of the main keywords characterizing this thesis.

Chapter 2

Sparse vectors

2.1 Introduction

Probably the most important and popular low-dimensional signal models are the *sparse signal models*. Some of the basic elements, results and tools concerning these models were developed several decades back, however the field witnessed a research explosion in 2004 that continues to present day. The attention the field received recently is due to the contributions of Candes, Tao [CT05] and Donoho [Don06], around the middle of the previous decade. The study of sparse signals led to a revolution in signal processing and data sciences, that produced new techniques, ideas and tools as well as solutions in many problems that could not be addressed previously.

The study of sparse signals is an interdisciplinary field of research, as it attracts scientists from different communities, such as machine learning, signal processing, imaging, statistics and applied mathematics. Also, it has proven very useful in practice as many application arose in different fields, including but not limited to computational biology, image processing, machine learning, telecommunications, radars, medical imaging, etc. The study of this field of research requires a solid theoretical background in linear algebra, probability theory, optimization theory, functional analysis and other fields depending on the specific problem.

2.2 Sparse and compressible signals

We are going to begin with the definition of a *sparse signal*.

Definition 2.2.1 (Sparse signal)**.** *A vector $\boldsymbol{x} \in \mathbb{R}^n$ is called k-sparse if it's non-zero entries are at most k, i.e $card(supp(\boldsymbol{x})) \leq k$.*

The set of all k-sparse signal is denoted as Σ_k. Also note that this is an idealized model in the sense that there are no sparse signals in reality, only signals that are approximately sparse. We are going to define an extension of sparse signal models that describes approximately sparse signals.

From a geometric point of view the set of all k-sparse signals is a union of subspaces. The k-sparse signals model accounts for all the $\binom{n}{k}$ different ways to pick the k non-zero elements of the arbitrary k-sparse vector. Every different

way defines a subspace spanned by the basis vectors corresponding to the chosen indices. Notice that each subspace covers the case that the vector possess at most k non-zero elements in the respective positions. Consequently, taking into account all the possible ways we end up with a union of subspaces.

In order to capture the model of signals that are approximately sparse we introduce *compressible signal models*. Before we proceed with the definition of a compressible signal we must define the l_p error of the best k-term estimate of a signal vector.

Definition 2.2.2 (l_p error of the best k-term estimate)**.** *[FR13] The l_p error of the best k-term estimate of a signal vector $\boldsymbol{x} \in \mathbb{R}^n$, with $p > 0$, is given by*

$$\sigma_k(\boldsymbol{x})_p := \underset{\boldsymbol{z} \in \Sigma_k}{inf} \|\boldsymbol{x} - \boldsymbol{z}\|_p.$$

It turns out that the best k-term estimate of an arbitrary vector $\boldsymbol{x} \in \mathbb{R}^n$ is simply the vector that is created by choosing the k larger elements of the original vector $\boldsymbol{x}$. Assuming that we have arranged the elements of the vector $\boldsymbol{x}$ in descending order, for the l_p error of the best k-term estimate of $\boldsymbol{x}$, we get obtain [Boc+15]

$$\sigma_k(\boldsymbol{x})_p = \begin{cases} \left(\sum_{i=k+1}^{n} |x_i|^p \right)^{1/p} & , p \in (0, +\infty) \\ x_{k+1} & , p = +\infty \end{cases}.$$

Now we can define a compressible signal vector. Roughly, a signal vector $\boldsymbol{x} \in \mathbb{R}^n$ is called *compressible* if the l_p error of its best k-term estimate, for some $p > 0$, decays quickly with respect to k. More formally we have the following definition.

Definition 2.2.3 (Compressible signal)**.** *[BCW10] A signal vector $\boldsymbol{x} \in \mathbb{R}^n$ is called compressible, if there exists $R > 0$ and $p \leq 1$ such that it holds*

$$|x_i| \leq R i^{-1/p}, \; 1 \leq i \leq n,$$

where we have assumed that the components were sorted in descending order.

Essentially the set of compressible signals is the set of signals that can be successfully approximated by sparse ones. There are several examples of sparse/ compressible signals. One of the most characteristic example is provided below.

Many real-world images are (approximately) sparse in specific bases, such as the Direct Cosine Transform (DCT) basis or the wavelet basis. This attribute of images is exploited by the JPEG-2000 standard. Assuming that we represent an image as a vector (by putting each column of the image's matrix sequentially), with each entry describing the grey-level intensity, the JPEG-2000 standard transforms the image vector to the wavelet domain in which the signal contains a small number of large coefficients. The final image is acquired by keeping those large coefficients and putting the rest to zero. As a result, there is a significant gain in storage space with only a small deterioration in the image's quality.

2.3 Signal dictionaries

In our exposition a signal is a vector $\boldsymbol{x} \in \mathbb{R}^n$ in the vector space $\mathbb{R}^n$. As a result, $\mathbb{R}^n$ admits a basis, in which $\boldsymbol{x}$ can be written in a unique way. In other words, every signal vector $\boldsymbol{x}$ can be written as a linear combination of the independent elements of a basis. There are many bases which can be used to describe a given vector space. In this thesis the vectors represent signals, so we are going to introduce some extra terms. In the nomenclature of signal processing the generalization of the basis is called a *dictionary* and the elements of the basis/ dictionary are called *atoms*. [The15]

Definition 2.3.1 (Dictionary)**.** *A dictionary is a set of elementary signal vectors $\boldsymbol{\psi}_i \in \mathbb{R}^n$, $i \in I$, not necessarily independent, that span a signal space. The elements of the dictionary are called atoms.*

The dictionaries can be characterized as *complete* and *overcomplete*. The definition of these terms follows.

Definition 2.3.2 (Complete and overcomplete dictionaries)**.**

- *A dictionary is called complete if it consists of n (usually orthonormal) independent signal vectors $\boldsymbol{\psi}_i \in \mathbb{R}^n$, $1 \leq i \leq n$. An arbitrary signal $\boldsymbol{x}$ can be written using the atoms of a complete dictionary in an unique way.*
- *A dictionary is called overcomplete if it consists of more than n signal vectors $\boldsymbol{\psi}_i \in \mathbb{R}^n$, $i \in I$, i.e more vectors than the dimension of the underlying vector space. The atoms are dependent, as their number is greater than the dimensionality of the underlying vector space. There are many ways to express an arbitrary signal $\boldsymbol{x}$ in such dictionaries.*

We can express a signal $\boldsymbol{x} \in \mathbb{R}^n$ in a given (overcomplete) dictionary Ψ as

$$\boldsymbol{x} = \Psi\boldsymbol{\theta}$$

or

$$\boldsymbol{x} = \sum_{i \in I} \theta_i \boldsymbol{\psi}_i,$$

where $\{\theta_i\}, i \in I$ are the coefficients of the signal in the dictionary Ψ and $\{\boldsymbol{\psi}_i\}, i \in I$ (the columns of matrix Ψ) the atoms of the dictionary. In general there is no unique set of coefficients $\{\theta_i\}, i \in I$ when we use an overcomplete dictionary. While it might seem that the existence of overcomplete dictionaries is mathematically flawed, since it abuses the properties of a vector space basis, there is a mathematical concept that describes these kind of dictionaries. This generalization of a basis is called a *frame*. [KC07]

Probably the most characteristic example is the Discrete Fourier base; the base that corresponds to the Discrete Fourier transform (DFT). If we put the DFT basis vectors as columns in a matrix we formulate the DFT matrix Ψ. Notice that Ψ is an orthogonal matrix as DFT is an orthogonal transformation, i.e its basis vectors are orthogonal to one another.

Given a set of signal vectors, $\{\boldsymbol{x}_i\}_{1 \leq i \leq s}$, we can formulate the problem of finding an (overcomplete) dictionary in which all the vectors admit a sparse representation. This task is called *dictionary learning* and it has the potential to enhance performance in many applications.

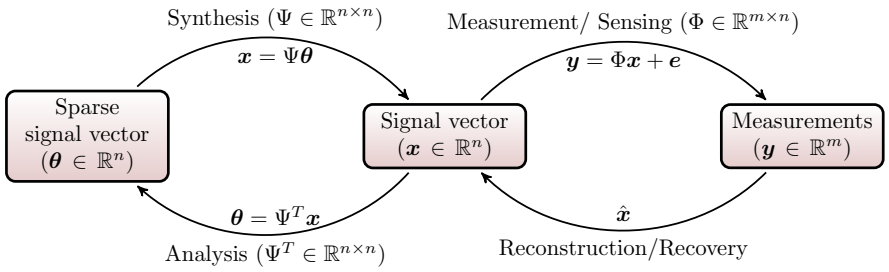

Figure 2.1: The big picture of the sparse vector recovery problem.

2.4 The big picture of sparse recovery

In this section we are going to give an outline of the setting (i.e the big picture) of the sparse recovery problem and define the basic elements that comprise it. This presentation aims at collecting the basic concepts and terms in one place. Figure 2.1 contains the basic elements and operations of the problem aside with the necessary terminology. Do not take this figure too literally or stick to the exact notation and meaning, as some things may be used differently at a later stage.

The signals depicted in figure 2.1 are the following.

- Signal vector ($\boldsymbol{x} \in \mathbb{R}^n$)
  The unknown signal vector, $\boldsymbol{x} \in \mathbb{R}^n$, that we observe through a linear measurement process. It is the signal we want to (implicitly, through estimation of $\boldsymbol{\theta}$) estimate based on the set of measurements $\boldsymbol{y}$ and exploiting the fact that there exists a basis Ψ in which $\boldsymbol{x}$ is sparse. It belongs to a high dimensional signal space (with respect to the dimension m of the measurements space) and we assume that is given in the canonical basis, in which the signal is not necessarily sparse.

- Observations/Measurements/Samples ($\boldsymbol{y} \in \mathbb{R}^m$)
  The observations vector $\boldsymbol{y} \in \mathbb{R}^m$ that results from sensing $\boldsymbol{x}$ with a measurement matrix Φ. It is a vector residing in a low dimensional space (with respect to the dimension n of the signal vector $\boldsymbol{x}$).

- Sparse signal vector ($\boldsymbol{\theta} \in \mathbb{R}^n$)
  The signal vector $\boldsymbol{x} \in \mathbb{R}^n$ expressed in a basis where it is sparse. It is the signal that we want to estimate given the observations $\boldsymbol{y}$ and the knowledge that $\boldsymbol{x}$ admits a sparsifying basis, in which it is represented as $\boldsymbol{\theta}$.

The basic operations we present are the following.

- Change of basis/Transformation (Ψ, Ψ^T)
  These matrices represent the change of basis matrices for the signal vectors $\boldsymbol{x}$ and $\boldsymbol{\theta}$. For simplicity we assume that the dictionary used is complete and orthogonal and so the transformations are represented by orthogonal matrices. As a result, it holds that $\Psi^{-1} = \Psi^T$ and consequently we have that $\boldsymbol{x} = \Psi\boldsymbol{\theta}$ (Synthesis) and $\boldsymbol{\theta} = \Psi^T\boldsymbol{x}$ (Analysis).

- Sensing/Measurement/Sampling/Projection/Dimensionality reduction (Φ)
  The sensing matrix Φ that is used to perform the linear measurements process[1]. The operation of measurement is $\boldsymbol{y} = \Phi\boldsymbol{x}$ in the noiseless case and $\boldsymbol{y} = \Phi\boldsymbol{x} + \boldsymbol{e}$ in the noisy one, where the vector $\boldsymbol{e} \in \mathbb{R}^m$ represents the noise.

- Recovery/Reconstruction
  Recovery constitutes the inverse operation of sensing. This process aims at estimating successfully the original signal vector, given the observations $\boldsymbol{y}$ and the prior knowledge that the signal vector is sparse in some basis.

Now that we have described the basic components of the problem we are dealing with we are going to formulate the corresponding mathematical problem. In the noiseless case the following formulas hold

$$\boldsymbol{x} = \Psi\boldsymbol{\theta}, \tag{2.1}$$

$$\boldsymbol{y} = \Phi\boldsymbol{x}. \tag{2.2}$$

Combining the equations (2.1), (2.2) we get

$$\boldsymbol{y} = \Phi\boldsymbol{x} = \Phi\Psi\boldsymbol{\theta} \Rightarrow \boldsymbol{y} = A\boldsymbol{\theta},$$

where $A = \Phi\Psi$.

The problem we are dealing with is the *sparse vector recovery problem* and can be stated as follows: We are given a signal vector $\boldsymbol{x}_0 \in \mathbb{R}^n$, which is k-sparse in some basis $\Psi \in \mathbb{R}^{n\times n}$ and can be represented as $\boldsymbol{\theta}_0 \in \Sigma_k$ on it, and a measurement matrix $\Phi \in \mathbb{R}^{m\times n}$. We perform a measurement that is described as follows, depending on the presence of noise:

- $\boldsymbol{y} = \Phi\Psi\boldsymbol{\theta}_0 = A\boldsymbol{\theta}_0$ (noiseless case)
- $\boldsymbol{y} = \Phi\Psi\boldsymbol{\theta}_0 + \boldsymbol{e} = A\boldsymbol{\theta}_0 + \boldsymbol{e}$, $\|\boldsymbol{e}\|_2 \leq \epsilon$ (noisy case)

The problem is, given A and $\boldsymbol{y}$, to successfully reconstruct the sparse signal vector $\boldsymbol{\theta}_0$ and consequently $\boldsymbol{x}_0$. The exact meaning of success depends on the type of scenario we are dealing with. As a result, we have the following recovery concepts, where we denote the recovered vector (the estimate) as $\hat{\boldsymbol{\theta}}$:

- *Exact recovery*, if $\hat{\boldsymbol{\theta}} = \boldsymbol{\theta}_0$, in the noiseless case
- *Robust recovery*, if $\|\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0\|$ is small (for some norm), in the noisy case. Roughly, we want the error to scale moderately with respect to the noise level ϵ.

Also, another extension of both scenarios, the noiseless and the noisy one, that we are going to consider refers to the possibility that the signal vector $\boldsymbol{x}_0$ we measure and we want to reconstruct is only approximately sparse (compressible). In that case we care about the *stability* of the solution $\hat{\boldsymbol{\theta}}$, i.e the estimation error

[1] Essentially every observation is a linear combination of the elements of the unknown signal vector, expressed as the inner product of the signal vector with some row of the measurement matrix, i.e $y_i = \boldsymbol{\phi}_i^T\boldsymbol{\theta}$, $i = 1, \ldots, m$. That is the reason we refer to sensing as a linear measurement process.

$\|\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0\|$ to scale slowly with respect to the approximation error $\sigma_k(\boldsymbol{\theta})_p$ of the compressible vector by a sparse one. In order to avoid confusion we are going to postpone the analysis of the noisy case and the discussion about stability issues.

Therefore, in the noiseless case we end up with the following underdetermined system of linear equations, whose solutions we seek.

$$\boldsymbol{y} = A\boldsymbol{\theta},\ A \in \mathbb{R}^{m\times n},\ \boldsymbol{y} \in \mathbb{R}^m,\ \boldsymbol{\theta} \in \mathbb{R}^n,\ m < n.$$

This system is underdetermined and as a result it admits an infinite number of solutions. The fact that allows us to overcome this obstacle is the prior knowledge that $\boldsymbol{x}$ is sparse in some basis. It is possible that the system, under certain conditions, admits a unique sparse solution, the original sparse signal $\boldsymbol{\theta}$. We are going to study this possibility, as well as the specific conditions in the next sections.

2.5 Underdetermined system of linear equations

As we mentioned previously the mathematical problem that lies behind the sparse signal recovery problem is the solution of an *underdetermined system of linear equations*, given the constraint that the solution we seek is sparse. First of all, we are going to study the general solution of the unconstrained problem. Therefore, we have [The15]

$$\boldsymbol{y} = A\boldsymbol{\theta},\ A \in \mathbb{R}^{m\times n},\ \boldsymbol{y} \in \mathbb{R}^m,\ \boldsymbol{\theta} \in \mathbb{R}^n,\ m < n. \tag{2.3}$$

Without loss of generality we assume that matrix A has full row rank, so it describes m independent linear equations of n variables. Obviously the set of solutions is infinite. Specifically, each equation defines a hyperplane in the n-dimensional space. Notice that the hyperplanes are non-parallel as a result of the linear independence of the respective equations. So the solutions set is created by the intersection of the m hyperplanes in the n-dimensional space. As a result, the intersection is a $(n-m)$-dimensional hyperplane. We denote the solution set as

$$\Theta = \left\{\boldsymbol{\theta} \in \mathbb{R}^n :\ y_i = \boldsymbol{a}_i^T\boldsymbol{\theta},\ i = 1,\dots,m\right\},$$

where $\boldsymbol{a}_i^T$ denotes the rows of A.

Equivalently this set can be written as

$$\Theta = \boldsymbol{\theta}_0 + null(A),$$

where $\boldsymbol{\theta}_0$ is a solution of (2.3). We can easily see that if we consider a fixed solution $\boldsymbol{\theta}_0$ and an arbitrary solution $\boldsymbol{\theta}$ of system (2.3). Then we have that

$$A\left(\boldsymbol{\theta}_0 - \boldsymbol{\theta}\right) = 0 \Rightarrow \boldsymbol{\theta}_0 - \boldsymbol{\theta} \in Null(A) \Rightarrow \boldsymbol{\theta} = \boldsymbol{\theta}_0 + Null(A).$$

In order to obtain the solutions of the constrained system of linear equations we are going to employ our intuition in finding the correct solution strategy. Probably the first idea that comes in mind is to search in the set of all possible solutions of the unconstrained system and pick the solutions that have the smallest number of non-zero elements. In the following subsection we are going to formulate that idea as an optimization problem.

2.6 The optimization tasks of sparse recovery

We are going to reduce the problem of (noiseless) sparse recovery to the solution of an optimization task of the form

$$\begin{aligned} \underset{\boldsymbol{\theta}\in\mathbb{R}^n}{\text{minimize}} \quad & \|\boldsymbol{\theta}\|_p \\ \text{subject to} \quad & \boldsymbol{y} = A\boldsymbol{\theta}. \end{aligned},$$

where $\|\cdot\|_p$ is some l_p norm. Actually, we are going to evaluate three different optimization tasks, corresponding to $p = 0, 1, 2$.

The setting in which we study the optimization tasks is the following: Suppose that we have a signal vector $\boldsymbol{x}_0$, which admits a sparsifying representation $\boldsymbol{\theta}_0 \in \Sigma_k$. We measure the signal $\boldsymbol{x}_0$ and as a result we obtain the observations $\boldsymbol{y} = \Phi\boldsymbol{x}_0 = A\boldsymbol{\theta}_0$, where $A = \Psi\Phi$. At many points we are going to consider the optimization tasks as algorithms that admit as input the matrix A and the observations $\boldsymbol{y}$ and produce as output an estimate $\hat{\boldsymbol{\theta}}$ of the vector $\boldsymbol{\theta}_0$. Then we can obtain an estimate of the original signal vector, $\hat{\boldsymbol{x}} = \Psi\hat{\boldsymbol{\theta}}$. Also, note that we are going to treat matrix A as the measurement/sensing matrix that measures the sparse signal vector $\boldsymbol{\theta}_0$, although we described it differently in the previous section.

2.6.1 l_0 norm

The analysis in this part relies on the notion of the l_0 norm. The l_0 norm is defined as $\|\boldsymbol{\theta}\|_0 = card(supp(\boldsymbol{\theta}))$. However, it is important to stress that technically the l_0 norm is not valid, as it does not satisfy all the necessary conditions in order to constitute a norm. Nevertheless, we are going to use the term "l_0 norm" keeping though in mind this technicality.

A first idea in the direction of solving the problem is described by the following optimization task.

Definition 2.6.1 (l_0 minimization task)**.**

$$\begin{aligned} \underset{\boldsymbol{\theta}\in\mathbb{R}^n}{minimize} \quad & \|\boldsymbol{\theta}\|_0 \\ subject\ to \quad & \boldsymbol{y} = A\boldsymbol{\theta}. \end{aligned}$$

Essentially, what the l_0 minimization task does is from the set of all possible solutions of the system of linear equations $\boldsymbol{y} = A\boldsymbol{\theta}$ returns the solutions that have the smallest l_0 norm, i.e the sparsest ones. Clearly we hope that there is only one solution that exhibits the smallest l_0 norm, the one that was sensed ($\boldsymbol{\theta}_0$) and its respective measurement $\boldsymbol{y} = A\boldsymbol{\theta}_0$ was given as input along with matrix A.

This problem is NP-hard in the general case, where we consider as possible input all the matrices $A \in \mathbb{R}^{m\times n}$ and observations vectors $\boldsymbol{y} \in \mathbb{R}^m$. The important issue here is whether the output of the l_0 minimization task returns a unique and correct sparse solution in all possible cases. The answer is no, since problems can occur if, for example, we pick the measurement matrix A in a way that will project (at least) two sparse signals (where one of them is the true signal; the one that was sensed and we seek to reconstruct) to the same vector in the low dimensional space.

More formally we have the following theorem that characterizes the adequacy of l_0 optimization task in recovering sparse vectors from linear measurements. [FR13]

Theorem 2.6.1. *Let $A \in \mathbb{R}^{m\times n}$ be a matrix. Then the following statements are equivalent.*

1. *For every $\boldsymbol{\theta}_0 \in \Sigma_k$, it holds that $\boldsymbol{\theta}_0$ is the unique k-sparse solution of the linear system $\boldsymbol{y} = A\boldsymbol{\theta}$, where $\boldsymbol{y} = A\boldsymbol{\theta}_0$ and $\boldsymbol{\theta} \in \mathbb{R}^n$.*

2. *For every $\boldsymbol{\theta}_0 \in \Sigma_k$, it holds that $\boldsymbol{\theta}_0$ is the unique solution of the l_0 minimization task*

$$\begin{aligned} &\underset{\boldsymbol{\theta}\in\mathbb{R}^n}{\text{minimize}} && \|\boldsymbol{\theta}\|_0 \\ &\text{subject to} && \boldsymbol{y} = A\boldsymbol{\theta} \end{aligned},$$

   where $\boldsymbol{y} = A\boldsymbol{\theta}_0$.

3. *For the matrix $A \in \mathbb{R}^{m\times n}$ it holds that*

$$KerA \cap \Sigma_{2k} = \{0\}.$$

Essentially, the previous statement provides a necessary and sufficient condition for a measurement matrix A that ensures that every k-sparse vector can be successfully reconstructed using (2.6.1), after sensing using A. Also, it says that if the system of equations has a unique k-sparse solution then the l_0 minimization task will find it. Finally, it is worth noting that for the number of measurements m that are necessary in order to reconstruct every k-sparse vector, using the same matrix, we can deduce that (at least in principle) $m \geq 2k$ observations are sufficient.

Consequently, regarding its ability to provide a unique and correct sparse solution we notice that this task captures the notion of finding the sparsest solution, but fails to deliver the solution we seek in all cases. It only works under certain conditions, depending on the choice of the measurement matrix A. Nevertheless, even in the case where the conditions will proven to be relaxed enough for us to claim that the correct solution is obtained in all practical situations, the computational intractability renders this task useless. As a result, we need to resort to a different approach.

2.6.2 l_2 norm

One idea in order to mitigate the intractability of the previous optimization task is to relax our objectives, deviating from the task that captures the notion of finding sparse vectors, hoping though that in many (practical) cases the new task will provide the correct solution. One such relaxation is contained in the following optimization task.

Definition 2.6.2 (l_2 minimization task)**.**

$$\begin{aligned} &\underset{\boldsymbol{\theta}\in\mathbb{R}^n}{\text{minimize}} && \|\boldsymbol{\theta}\|_2 \\ &\text{subject to} && \boldsymbol{y} = A\boldsymbol{\theta}. \end{aligned}$$

The solution of this problem is unique and is given by the closed form expression [The15]

$$\boldsymbol{\theta} = A^T(AA^T)^{-1}\boldsymbol{y}.$$

It is straightforward that this problem is computationally feasible as computing its solution requires only matrix multiplications and inversions, tasks for which we possess efficient algorithms. Again the important question is whether the l_2 minimization task provides a unique and correct sparse solution. Although a unique solution is guaranteed, this solution is not necessarily the correct one, the one that was sensed. Actually in many cases the l_2 minimization task it doesn't even return sparse solutions. Hence, the l_2 optimization task does not have practical value for our problem and so we are going to omit a more detailed analysis.

Consequently, this task, although computationally tractable, it does not always provides us with sparse solutions. So we are forced to consider another relaxation.

2.6.3 l_1 norm

The middle ground between the previous two optimization tasks is the l_1 *minimization task.*

Definition 2.6.3 (l_1 minimization task)**.**

$$\begin{aligned} &\underset{\boldsymbol{\theta}\in\mathbb{R}^n}{\textit{minimize}} && \|\boldsymbol{\theta}\|_1 \\ &\textit{subject to} && \boldsymbol{y} = A\boldsymbol{\theta}. \end{aligned}$$

This problem can be reduced to a linear programming task, which is a problem that belongs to complexity class P. Thus, there are several efficient algorithms which can be used to obtain the solutions of such tasks (e.g interior point methods, simplex methods).

The main question here is whether the l_1 minimization task can provide us the correct sparse solution. As anyone by now might have anticipated the answer is: under certain conditions, depending on the choice of the sensing matrix A. The different outcomes of the l_1 minimization task are the following

- An infinite number of solutions
- A unique sparse solution, but not the correct one
- The correct sparse solution (unique solution).

The different outcomes highlight the importance of choosing the proper sensing matrix. Figure 2.2 illustrates these three different cases. Next, we are going to present a necessary and sufficient condition for matrix A, such that the l_1 minimization task returns the correct sparse solution, for every possible k-sparse vector $\boldsymbol{\theta} \in \mathbb{R}^n$ given as input. The condition we will discuss is called the *null-space property.*

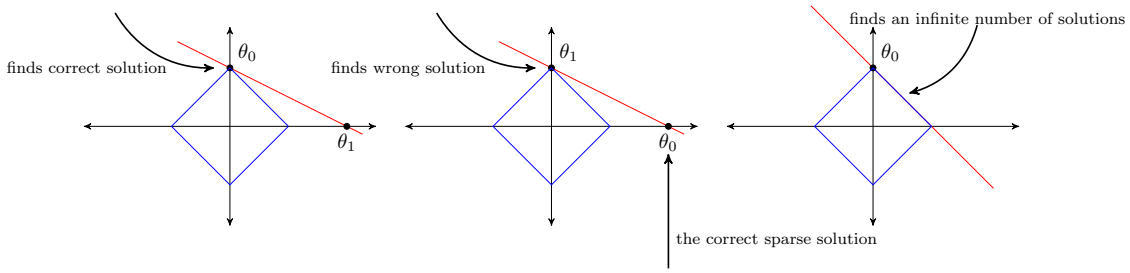

Figure 2.2: The three different scenarios that can arise in the sparse vector recovery problem using the l_1 minimization task. Suppose that we have $n = 2$, $m = 1$, $\boldsymbol{\theta}_0$ is the 1-sparse vector we sense and $\boldsymbol{\theta}_1$ some other 1-sparse vector. (a) The l_1 minimization task finds the correct sparse solution. (b) The l_1 minimization task finds a 1-sparse solution but not the correct one ($\boldsymbol{\theta}_0$). (c) The l_1 minimization task returns an infinite number of solutions including the correct one. It is obvious that only case (a) is considered successful. The figure is based on [The15].

Definition 2.6.4 (Null-space property of order k)**.** *[FR13] Let $A \in \mathbb{R}^{m \times n}$ be a matrix. We say that A satisfies the null space property of order k if it holds that*

$$\|\boldsymbol{x}_S\|_1 < \|\boldsymbol{x}_{[n]\setminus S}\|_1, \forall \boldsymbol{x} \in KerA \setminus \{0\} \text{ and } \forall S \subseteq [n] \text{ with } card(S) \leq k, \quad (2.4)$$

where $[n] = \{1, 2, \ldots, n\}$.

Now we can state the following theorem that links the null space property with exact recovery of sparse vectors using the l_1 minimization task.

Theorem 2.6.2 (Null-space property and exact recovery)**.** *[Cha+12a] Let $A \in \mathbb{R}^{m \times n}$ be a matrix. The following statements are equivalent*

1. *For every $\boldsymbol{\theta}_0 \in \Sigma_k$, it holds that $\boldsymbol{\theta}_0$ is the unique solution of the l_1 minimization task, with $\boldsymbol{y} = A\boldsymbol{\theta}_0$.*

2. *The null space property of order k holds for matrix A.*

2.7 Compressed sensing

One possible way to exploit the fact that a signal is sparse in some basis is in data compression schemes. To illustrate that we consider the following example. Suppose that we have a signal $\boldsymbol{x} \in \mathbb{R}^n$ that is (approximately) sparse in some basis, where it is expressed as $\boldsymbol{\theta} \in \mathbb{R}^n$. Essentially, that means that we have performed a sampling procedure before that resulted in n observations (the n elements of vector $\boldsymbol{x}$). We want to exploit the signal's sparsity in order to compress it. We use the following steps in order to do that. [The15]

1. Transform the signal to the basis where it admits an approximate k-sparse representation. That is performed by $\boldsymbol{\theta} = \Phi\boldsymbol{x}$.

2. Keep the k largest values of $\boldsymbol{\theta}$ and encode their locations and their values. We can store a compressed version of the original signal using the previous encoding.

3. Create the signal vector $\boldsymbol{\theta}_0$, where the vector $\boldsymbol{\theta}_0$ is equal to $\boldsymbol{\theta}$ in the positions that correspond to the k largest elements of $\boldsymbol{\theta}$ and 0 in the rest. Using the inverse transform $\boldsymbol{x}_0 = \Phi^T \boldsymbol{\theta}_0$ we can obtain an approximate version $\boldsymbol{x}_0$ of the original signal $\boldsymbol{x} \in \mathbb{R}^n$, when this is necessary.

The previous procedure appears to be effective for compressing (approximately sparse) signals, but there is one thing that needs further consideration. We sample/measure n elements of a vector $\boldsymbol{x} \in \mathbb{R}^n$, but in the end we keep only k of the coefficients, where usually $k << n$. That makes us wonder whether we can do better. What if we managed to sample less than n elements of the n-dimensional signal $\boldsymbol{x}$, just enough in order to provide us with the k coefficients that adequately describe the signal in the sparsifying basis? That is the motivating idea that gave birth to the field of *Compressed sensing* or *Compressive sampling*. Ideally we would like to obtain a number of samples m close to $k << n$ (obviously $k < m$ must hold). In other words, we directly obtain a reduced number of samples, merging the steps of compression and sampling, that is the lowest possible, such that the information content of the signal (contained in the k largest coefficients) is preserved.

Compressed sensing is essentially a technique in signal processing. Leaving for a moment aside the exact formulation of the sparse recovery problem we used until now and considering it in a more general framework we have the following characteristic example that we should not omit from our exposition. We know from Nyquist's sampling theorem that in order to be able to perfectly reconstruct a bandlimited (continuous-time) signal we must sample it with sampling frequency $f_s \geq 2f$, where f_s is the sampling frequency and f is the highest frequency of the signal. We call the smallest possible sampling frequency, i.e $f_N = 2f$, *Nyquist frequency*. The field of compressed sensing essentially claims that given that the signal is sparse in some (continuous) basis (e.g Fourier basis) we can sample it with frequency lower that the Nyquist frequency and be able to perfectly reconstruct it. This capability is important in cases where the signals are wideband.

2.8 Measurement matrices

One of the most important aspects of the field of sparse recovery is that of designing the suitable *measurement matrix*. It is crucial to design the proper measurement matrix as it is the main factor that determines whether we can recover the correct sparse solution. Essentially, what we need is a measurement matrix that provides a stable embedding for the set of k-sparse vectors. In order to evaluate the capability of an arbitrary measurement matrix in performing a stable embedding we introduce two measures. The two measures are given below.

1. Coherence

2. Restricted isometry property

2.8.1 Coherence

A measure that characterizes the ability of a matrix to provide a stable embedding for the class of k-sparse vectors is *mutual coherence*. The definition is given below.

Definition 2.8.1 (Mutual coherence)**.** *[The15] Let $A \in \mathbb{R}^{m\times n}$ be an $m \times n$ matrix and $\boldsymbol{a}_i$, $1 \leq i \leq n$ be the columns of A. The mutual coherence of A is defined as*

$$\mu(A) = \max_{1\leq i<j\leq n} \frac{\left|\boldsymbol{a}_i^T \boldsymbol{a}_j\right|}{\|\boldsymbol{a}_i\|_2 \|\boldsymbol{a}_j\|_2}.$$

Roughly, mutual coherence is a measure of orthogonality of the columns of a matrix A. In other words, mutual coherence measures the "correlation" or "independence" between the columns of a matrix. In an orthogonal matrix A, where the columns are orthogonal, the mutual coherence is $\mu(A) = 0$. In non-square matrices with $(n > m)$, where full orthogonality is impossible, mutual coherence shows how close a matrix is to an orthogonal one. Specifically for the mutual coherence of an arbitrary matrix A it holds that $0 \leq \mu(A) \leq 1$. If $n > m$ then we can obtain an improved result, known as Welch bound,

$$\sqrt{\frac{n-m}{m(n-1)}} \leq \mu(A) \leq 1.$$

In the context of sparse recovery, the smaller the mutual coherence of a matrix A, the better, in the sense of matrix A is able to handle greater levels of sparsity and recovery becomes easier. Intuitively, we want the columns of the measurement matrix to be as "independent/uncorrelated" as possible (small mutual coherence). Then in the formulation of the observations vector [2], each component of the sparse signal vector is provided by a column, which has small "correlation" with the other columns. Roughly that means that the information encompassed in each component of $\boldsymbol{\theta}$ is captured by a column that is "uncorrelated" ("independent") with the other columns, hence making the process of separating these components and unveiling the information they contain easier. As a result, a significant amount of the information content of the original signal is preserved and recovery becomes easier.

2.8.2 Restricted isometry property

The *Restricted Isometry Property (R.I.P)* is probably the most important measure.

Definition 2.8.2 (Restricted Isometry Property (R.I.P) condition)**.** *[The15] Let δ_k, $1 \leq k \leq n$ be the kth restricted isometry constant of a matrix $A \in \mathbb{R}^{m\times n}$. It is defined as the smallest constant such that the following condition holds*

$$(1-\delta_k)\|\boldsymbol{\theta}\|_2^2 \leq \|A\boldsymbol{\theta}\|_2^2 \leq (1+\delta_k)\|\boldsymbol{\theta}\|_2^2, \ \forall \boldsymbol{\theta} \in \Sigma_k. \tag{2.6}$$

[2] Note that the operation $A\boldsymbol{\theta}$ can be equivalently written as

$$\boldsymbol{y} = A\boldsymbol{\theta} = \sum_{i=1}^{n} \theta_i \boldsymbol{a}_i, \tag{2.5}$$

where $\boldsymbol{a}_i$, $1 \leq i \leq n$ are the columns of A.

Roughly, we say that the Restricted Isometry Property of order k holds if equation 2.6 holds and δ_k is sufficiently smaller than one. Essentially, if a measurement matrix satisfies the R.I.P that means that the l_2 norm of the projection of every k-sparse vector is approximately preserved. Also, note that we want the measurement matrix A to satisfy the R.I.P with as high k as possible, since that means that it is capable of providing information preserving projections for a wider range of sparsity levels.

It is easy to see that if the measurement matrix A manages to provide a projection with a certain accuracy for some sparsity level k, it can also provide at least the same accuracy to vectors with smaller levels of sparsity. Hence, for the R.I.P constants it holds that

$$\delta_1 \leq \delta_2 \leq \cdots \leq \delta_k \leq \delta_{k+1} \leq \cdots \leq \delta_n. \tag{2.7}$$

We also consider another form of the R.I.P. Let $\boldsymbol{\theta}_1$, $\boldsymbol{\theta}_2 \in \Sigma_k$ be two k-sparse vectors. Then, $\boldsymbol{\theta}_1 - \boldsymbol{\theta}_2$ is a $2k$-sparse vector, so the R.I.P condition takes the following form.

$$(1-\delta_{2k})\|\boldsymbol{\theta}_1 - \boldsymbol{\theta}_2\|_2^2 \leq \|A(\boldsymbol{\theta}_1 - \boldsymbol{\theta}_2)\|_2^2 \leq (1+\delta_{2k})\|\boldsymbol{\theta}_1 - \boldsymbol{\theta}_2\|_2^2. \tag{2.8}$$

If a measurement matrix satisfies the R.I.P of order $2k$, i.e expression (2.6) holds with δ_{2k} sufficiently small, then the expression (2.8) dictates that the l_2 distances between every pair of k-sparse vectors are approximately preserved after the projection in the lower dimensional space. Thus, we can see that when the R.I.P holds it is straightforward that the projection provided by A is a stable embedding for the class of k-sparse vectors.

Matrices that obey the R.I.P

Establishing the R.I.P for a general matrix is a difficult task. Therefore, we need to find classes of matrices for which the R.I.P can be evaluated in an efficient way. Perhaps the most important class of matrices that obey, under certain conditions, the R.I.P are specific classes of random matrices.

Several characteristic classes of random matrices for which, under certain conditions, the R.I.P holds with high probability are the following. [FR13] [CW08]

- Gaussian random matrices
  A (normalized) Gaussian random matrix is a matrix $A \in \mathbb{R}^{m\times n}$, whose entries are i.i.d Gaussian random variables with mean $\mu = 0$ and variance $\sigma^2 = \frac{1}{m}$, i.e $A_{i,j} \sim \mathcal{N}(0, \frac{1}{m})$.
- Bernoulli random matrices
  A (normalized) Bernoulli random matrix is a matrix $A \in \mathbb{R}^{m\times n}$, whose entries are i.i.d Bernoulli random variables. The Bernoulli random variables in this case take the following values

$$A_{i,j} = \begin{cases} \frac{1}{\sqrt{m}} & \text{with probability} \quad 1/2 \\ -\frac{1}{\sqrt{m}} & \text{with probability} \quad 1/2 \end{cases}.$$

- Subgaussian random matrices
  A (normalized) subgaussian random matrix is a matrix $A \in \mathbb{R}^{m \times n}$, whose entries are i.i.d subgaussian random variables, i.e $\mathbb{P}\left[|A_{i,j}| > t\right] \leq e^{-t^2/2}, \forall t > 0$. Note that the previous two classes are special cases of subgaussian random matrices.

One might have noticed that in the scenario we described earlier we can freely choose the measurement matrix, but we obviously do not have a choice on the transformation matrix, since we cannot pick the basis in which the signal is sparse. As a result, we cannot completely specify matrix $A = \Phi\Psi$. At first this seems to be a problem as we need to specify the matrix Φ with respect to Ψ, in order for the R.I.P to hold for the matrix $A = \Phi\Psi$. However, in many cases, this is not a problem. For example, the classes of random matrices we mentioned before provide a universal choice of projection matrices that work for any orthonormal basis, i.e matrix $A = \Phi\Psi$ satisfies the R.I.P, when Φ is a suitable random matrix and Ψ is some orthogonal dictionary. Hence, we can choose the matrix Φ independently from the orthogonal basis matrix Ψ.

Random matrices offer a practical way to construct matrices that provably obey the R.I.P with high probability. Establishing that under certain conditions the R.I.P holds for subgaussian matrices is going to be one of the main topics of the second part of this thesis, so we postpone the detailed analysis of the above topic for the next chapters.

2.9 Reconstruction schemes

The are several types of algorithms/reconstruction schemes for sparse recovery. The most important ones are the following. [The15]

1. Greedy algorithms
2. Convex optimization algorithms
3. Iterative shrinkage algorithms

We are going to evaluate the following aspects of the sparse recovery algorithms. [FR13]

1. Computational complexity
   A very important aspect of an algorithm is its *computational complexity*, that is the amount of computational resources it requires in order to provide an output, with respect to input size and other parameters. Specifically we care about time complexity, i.e roughly the number of steps required in order to compute the solution, with respect to the input size.
2. Stability
   Stability concerns the ability of the reconstruction scheme to provide satisfactory solutions in cases where the signal vector is not exactly sparse, but it is compressible. It is logical to demand from an algorithm to work successfully in more realistic scenarios (remember that in reality there are no sparse signals), in the sense of introducing an estimation error that is

controlled by the approximation error of the compressible signal vector by a sparse one.

3. Robustness
   Robustness refers to the ability of an algorithm to tolerate errors introduced to the measurement process, either by noise or by our inability to measure a quantity with infinite precision. We expect that the estimate produced by a robust algorithm to deviate from the correct value by an amount that is controlled by the measurement error.

It is vital for the reconstruction algorithms to possess the last two qualities and at the same time to be able to work fast in order to have practical significance. In the opposite case the algorithm cannot be applied to practical situations and has purely theoretical significance. Taking into consideration more realistic scenarios we modify the formulation of the sparse recovery problem in order to accommodate for noise, measurement inaccuracies and compressible vectors.

For the noisy case we know from section 2.4 that $\boldsymbol{y} = A\boldsymbol{\theta} + \boldsymbol{e}$, $\|\boldsymbol{e}\|_2 \leq \epsilon$ and as result we have that

$$\|\boldsymbol{y} - A\boldsymbol{\theta}\|_2 \leq \epsilon \tag{2.9}$$

Therefore, we develop an extension of the l_1 minimization task (2.6.3), which we call *noisy or robust l_1 minimization task*. [The15]

Definition 2.9.1 (noisy l_1 minimization task)**.**

$$\begin{aligned} &\underset{\boldsymbol{\theta}\in\mathbb{R}^n}{\textit{minimize}} && \|\boldsymbol{\theta}\|_1 \\ &\textit{subject to} && \|\boldsymbol{y} - A\boldsymbol{\theta}\|_2 \leq \epsilon. \end{aligned} \tag{2.10}$$

We should note that in this case seeking a unique sparse solution no longer makes sense. Instead, the main issues here are robustness and stability (the last aspect was already an issue in the noiseless case). In other words, we expect from the algorithms that solve the previous task to be able to withstand noise and deviations from the sparse signal model, in the sense of introducing an estimation error that is roughly controlled by the noise level ϵ and the l_p error of the best k-term estimate $\sigma_k(\boldsymbol{\theta}_0)_p$ of the original signal $\boldsymbol{\theta}_0$ (the signal that was sensed).

We are going to study one typical example from each type of reconstruction schemes. For every reconstruction scheme we will provide a theorem concerning the estimation error of the respective scheme for arbitrary input, when given some guarantee about the R.I.P constant of the measurement matrix. The general formulation of that theorem would be as follows.

Theorem 2.9.1 (Sufficient condition for robust and stable recovery)**.** *[FR13] Suppose that the restricted isometry property of order rk holds for a matrix $A \in \mathbb{R}^{m\times n}$, with constant $\delta_{rk} < \delta < 1$. Then for any $\boldsymbol{\theta} \in \mathbb{R}^n$ an output $\hat{\boldsymbol{\theta}}$ of algorithm Alg, with input A, $\boldsymbol{y} = A\boldsymbol{\theta} + \boldsymbol{e}$ and ϵ satisfies the following error bounds*

$$\|\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}\|_1 \leq C_0\sigma_k(\boldsymbol{\theta})_1 + C_1\sqrt{k}\epsilon \tag{2.11}$$

and

$$\|\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}\|_2 \leq C_0 \frac{\sigma_k(\boldsymbol{\theta})_1}{\sqrt{k}} + C_1 \epsilon, \tag{2.12}$$

where C_0, C_1 are constants that depend only on δ_{rk}.

For every reconstruction scheme *Alg* we consider we provide the estimation error in the form given in the above theorem (with different constants C_0, C_1 for each scheme), for specific choices of the R.I.P constant (r and δ, different for every scheme). It is straightforward to see that the algorithm *Alg*, under the condition that the theorem 2.9.1 holds for a reasonable R.I.P constant ($\delta < 1$), is stable and robust as the estimation error scales moderately with respect to the approximation error $\sigma_k(\boldsymbol{\theta})_1$, the noise level ϵ, as well as the sparsity level k. As a result, the algorithms that satisfy theorem 2.9.1 with a reasonable R.I.P constant are considered successful.

2.9.1 Greedy algorithms

In general, a greedy algorithm is an algorithm that at each stage takes a locally optimal choice, which may not necessarily lead to globally optimal solution, although that is the final aim. A greedy algorithm for sparse reconstruction [TW10] iteratively improves the current estimate of $\boldsymbol{\theta}$, using locally optimal updates of its coefficients at each iteration, in order to reduce the estimation error. There are several greedy algorithms for sparse reconstruction. Some of them are the following

- Orthogonal matching pursuit (OMP)
- Least angle regression (LARS)
- Compressed sensing matching pursuit (CSMP)
  - CoSaMP
  - Subspace pursuit

The algorithm that we are going to study in more detail is *Orthogonal Matching Pursuit (OMP)*.

Orthogonal matching pursuit

Orthogonal Matching Pursuit is one of the oldest algorithms for sparse reconstruction. A detailed description of OMP is provided next. [The15]

Orthogonal Matching Pursuit (OMP)
Input : Measurement matrix $A \in \mathbb{R}^{m \times n}$, observations $\boldsymbol{y} \in \mathbb{R}^m$.
Output : A k-sparse estimate $\hat{\boldsymbol{\theta}} \in \mathbb{R}^n$.
Parameters :

- $\hat{\boldsymbol{\theta}}^{(i)}$: the i-sparse estimate at iteration i
- $\boldsymbol{e}^{(i)}$: the error at iteration i, i.e $\boldsymbol{e}^{(i)} = \boldsymbol{y} - A\hat{\boldsymbol{\theta}}^{(i)}$.
- $S^{(i)}$: the support set at iteration i

- $A^{(i)}$: the matrix formulated by considering the columns of A corresponding to the indices of $S^{(i)}$, i.e the active columns
- ϵ : the termination tolerance

Algorithm:

1. **Initialization**
   $\hat{\boldsymbol{\theta}}^{(0)} = \mathbf{0}$, $\boldsymbol{e}^{(0)} = \boldsymbol{y}$, $S^{(0)} = \emptyset$, $i = 1$.

2. **Identification of the column maximally correlated to error vector**
   Select the column a_{j_i} of A that exhibit the maximal correlation with the error vector ($\boldsymbol{e}^{(i)} = \boldsymbol{y} - A\hat{\boldsymbol{\theta}}^{(i)}$) at the previous iteration. Specifically, we pick the column
   $$j_i = arg \max_{j=1,\dots,n} \frac{\left|\boldsymbol{a}_j^T \boldsymbol{e}^{(i-1)}\right|}{\|\boldsymbol{a}_j\|_2}. \tag{2.13}$$

3. **Update of the support set**
   The support set at iteration i is
   $$S^{(i)} = S^{(i-1)} \cup \{j_i\}.$$

4. **Update of the sparse vector estimate**
   We can obtain the new estimate by solving the following least-squares problem
   $$\overline{\boldsymbol{\theta}} = arg \min_{\boldsymbol{z} \in \mathbb{R}^i} \|\boldsymbol{y} - A^{(i)}\boldsymbol{z}\|_2^2.$$
   Then we can obtain $\hat{\boldsymbol{\theta}}^{(i)}$ by taking the elements of $\overline{\boldsymbol{\theta}}$ and inserting them in the positions specified by the support set $S^{(i)}$, while setting the elements in the other positions to 0.

5. **Update of the error vector**
   The error at iteration i is
   $$\boldsymbol{e}^{(i)} = \boldsymbol{y} - A\hat{\boldsymbol{\theta}}^{(i)}.$$

6. **Condition check**
   The user picks a constant ϵ at the beginning as a termination tolerance. If $\boldsymbol{e}^{(i)} < \epsilon$ then the algorithm terminates, else the algorithm moves back to step 2.

Essentially, OMP is based on the idea that at iteration i the column which exhibits the maximum correlation (equation (2.13)) with the error vector $\boldsymbol{e}^{(i-1)}$ is the one that leads to the highest reduction of the l_2 norm of the error, when taking into account all the active columns in the formulation of the new error vector. Furthermore, notice that the error vector is orthogonal to the space spanned by the active columns, i.e $\boldsymbol{e}^{(i)} \perp \{\boldsymbol{x}_{j_1}, \dots, \boldsymbol{x}_{j_i}\}$, since $\hat{\boldsymbol{\theta}}^{(i)}$ is the optimal solution in the least squares sense. This property establishes that at the next

iteration the algorithm will not select a column that has already been chosen, i.e an active column. [The15]

Also, notice that the algorithm after k iterations returns a k-sparse solution. However, there is no guarantee that the final solution is a successful estimate of the correct one. The only thing we know is that the error (in the l_2 norm sense) decreases at every iteration. Theorem (2.9.2) essentially provides some conditions for matrix A under which OMP returns solutions with specific guarantees. Finally, the computational complexity of OMP is $\mathcal{O}(rnm)$, where r is the level of sparsity of the final solution. [The15]

We provide a theorem containing a sufficient condition for robust and stable recovery with OMP, as well as some performance bounds. [FR13]

Theorem 2.9.2 (Sufficient condition for robust and stable recovery with OMP)**.** *Consider the scenario described in theorem (2.9.1) and suppose that the restricted isometry property of order* $13k$ *holds for a matrix* $A \in \mathbb{R}^{m\times n}$ *with constant* $\delta_{13k} < 0.1666$ *(*$r = 13$, $\delta = 0.1666$*). Then for Orthogonal Matching Pursuit (OMP) theorem 2.9.1 is satisfied.*

2.9.2 Convex optimization algorithms

Convex optimization algorithms offer an attractive approach for sparse recovery. These algorithms solve convex optimization problems and therefore there are polynomial time algorithms that can be used (e.g interior point methods). The algorithm that we are going to discuss is *robust* l_1 *minimization.*

Robust l_1 minimization

Robust l_1 minimization is simply the algorithm that solves the robust variant of the familiar l_1 minimization task we introduced previously as robust l_1 minimization task.

Definition 2.9.2 (robust l_1 minimization)**.** *Let* $A \in \mathbb{R}^{m\times n}$ *be a measurement matrix and* ϵ *be the noise level. The optimization task robust* l_1 *minimization solves is*

$$\begin{aligned} \underset{\boldsymbol{\theta}\in\mathbb{R}^n}{\text{minimize}} \quad & \|\boldsymbol{\theta}\|_1 \\ \text{subject to} \quad & \|\boldsymbol{y} - A\boldsymbol{\theta}\|_2 \leq \epsilon. \end{aligned} \tag{2.14}$$

The following theorem contains a sufficient condition for robust and stable recovery with robust l_1 minimization. [FR13]

Theorem 2.9.3 (Sufficient condition for robust and stable recovery with robust l_1 minimization)**.** *Consider the scenario described in theorem (2.9.1) and suppose that the restricted isometry property of order* $2k$ *holds for a matrix* $A \in \mathbb{R}^{m\times n}$ *with constant* $\delta_{2k} < \dfrac{4}{\sqrt{41}} \approx 0.6246$ *(*$r = 2$, $\delta = 0.6246$*). Then robust* l_1 *minimization satisfies theorem 2.9.1.*

2.9.3 Iterative thresholding algorithms

This class of algorithms can be considered an extension of the classical iterative schemes for the solution of linear systems of equations, such as Gauss-Seidel

and Jacobi algorithms, to underdetermined systems of linear equations. The general formula of the iterative scheme is the following [The15]

$$\hat{\boldsymbol{\theta}}^{(i+1)} = F_i\left(\hat{\boldsymbol{\theta}}^{(i)} + Z(\boldsymbol{y} - A\hat{\boldsymbol{\theta}}^{(i)})\right), \tag{2.15}$$

for some matrix Z. The function F_i is a nonlinear thresholding operator that is applied component-wise. The two most notable choices for F_i is the *hard-thresholding operator* and the *soft-thresholding operator*. The hard-thresholding operator H_k acts on a vector and keeps its k largest elements unchanged, while putting the rest of the elements to zero. The soft-thresholding operator S_a sets to zero all the elements of the vector whose values are below the threshold a and reduces the magnitude of the rest by a. Some basic iterative thresholding algorithms are the following [FR13]

- Basic thresholding
- Iterative Hard Thresholding (IHT)
- Hard Thresholding Pursuit (HTP)

Iterative hard thresholding

Iterative hard thresholding (IHT) follows the formula (2.15) described before with $F_i = H_k$ and $Z = \delta A^T$, for some parameter δ which may depend on the iteration number. Roughly, choosing $Z = A^T$ makes sense since in sparse recovery scenarios we want the measurement matrix A to be as orthogonal as possible (low coherence) and therefore it holds that

$$\boldsymbol{y} = A\boldsymbol{\theta} \Rightarrow A^T\boldsymbol{y} = A^TA\boldsymbol{\theta} \approx \boldsymbol{\theta}.$$

A detailed description of IHT is provided next.

Input : Measurement matrix $A \in \mathbb{R}^{m\times n}$, observations $\boldsymbol{y} \in \mathbb{R}^m$, sparsity level k

Output : A k-sparse estimate $\hat{\boldsymbol{\theta}} \in \mathbb{R}^n$

Parameters :

- $\hat{\boldsymbol{\theta}}^{(i)}$: the estimate at iteration i
- ϵ : the termination tolerance

1. **Initialization**
   $\hat{\boldsymbol{\theta}}^{(0)} = \mathbf{0}$.
2. **Update of the estimate**
   $$\hat{\boldsymbol{\theta}}^{(i+1)} = H_k(\hat{\boldsymbol{\theta}}^{(i)} + A^T(\boldsymbol{y} - A\boldsymbol{\theta}^{(i)})),$$
   where H_k is the hard-thresholding operator.
3. **Condition check**
   If $\boldsymbol{y} - A\hat{\boldsymbol{\theta}}^{(i)} < \epsilon$ then the algorithm terminates, else the algorithm moves back to step 2.

The following theorem contains a sufficient condition for robust and stable recovery and an assessment of the estimation error of IHT. [FR13]

Theorem 2.9.4 (Sufficient condition for robust and stable recovery with IHT)**.** *Consider the scenario described in theorem (2.9.1) and suppose that the restricted isometry property of order* $3k$ *holds for a matrix* $A \in \mathbb{R}^{m \times n}$ *with constant* $\delta_{3k} < \frac{1}{\sqrt{3}} \approx 0.5774$ *(*$r = 3$*,* $\delta = 0.5774$*). Then Iterative Hard Thresholding (IHT) satisfies theorem 2.9.1.*

2.10 Applications

The development of the theoretical aspects of sparse signal models led to many practical applications. After all several of the tools used in this field were discovered many years before and their creation was motivated by practical applications. The list of applications includes geophysics, sampling, machine learning, radars, medical imaging, image processing, neuroscience, telecommunications, computer vision, etc.

The study of sparsity led to significant advances in the solution of linear inverse problems. We are going to give a characteristic example of how the knowledge of sparsity allows us to solve an inverse problem that was previously considered ill-posed. [The15] Suppose that we have a signal $\boldsymbol{x} \in \mathbb{R}^n$, which admits a sparse representation $\boldsymbol{\theta} \in \mathbb{R}^n$ in a basis Ψ. Signal $\boldsymbol{x}$ suffers some kind of distortion (e.g blurring in an image), a procedure described by a linear operator D. Also, we consider some noise $\boldsymbol{e}$ and consequently the final signal can be written as

$$\boldsymbol{y} = D\boldsymbol{x} + \boldsymbol{e} = D\Psi\boldsymbol{\theta} + \boldsymbol{e}. \tag{2.16}$$

Solving the corresponding inverse problem entails obtaining a good estimate $\hat{\boldsymbol{\theta}}$ of $\boldsymbol{\theta}$, given D, Ψ and $\boldsymbol{y}$. Then, we can estimate the original signal $\boldsymbol{x}$, as $\hat{\boldsymbol{x}} = \Psi\hat{\boldsymbol{\theta}}$. Exploiting the fact that the original signal is sparse in some basis we can formulate the solution of the inverse problem as a noisy l_1 minimization task, i.e

$$\begin{aligned} &\underset{\boldsymbol{\theta} \in \mathbb{R}^n}{\operatorname{argmin}} && \|\boldsymbol{\theta}\|_1 \\ &\text{subject to} && \|\boldsymbol{y} - D\Psi\boldsymbol{\theta}\|_2 \leq \epsilon. \end{aligned}$$

The previous example can describe a wide range of situations, such as image impainting, signal restoration, signal denoising, etc.

One of the most important applications of sparsity is in Magnetic Resonance Imaging (MRI). MRI is a medical imaging technology that is used in tasks such as brain imaging and angiography. It is able to produce pictures of the anatomy and the processes of the human body using radiowaves and magnetic fields. As a result, it does not expose the patients to harmful ionizing radiation. However, the time required to obtain a high-resolution picture (in other words perform a set of measurements) is generally high, ranging from several minutes to hours. Compressed sensing can be employed to overcome this shortcoming by requiring less samples and consequently reducing the time of an MRI scan.

Chapter 3

Low-rank matrices

3.1 Introduction

Low-rank matrices signal models are another important class of low-dimensional models that captured the interest of machine learning's community the recent years and is currently a very active topic of research. Here the main problem of interest is low-rank matrix recovery and especially a special case of this problem called *matrix completion*. In matrix completion problems we are given a matrix with missing entries and the objective is to complete those entries based on the prior knowledge that the underlying matrix is low-rank. Low-rank matrix recovery has numerous applications, including quantum mechanics, recommendation systems, sensor networks and among others the famous Netflix problem.

3.2 Preliminaries about matrices

The main object of interest in this chapter are low-rank matrices. Therefore, we need to introduce some basic definitions and results about matrices. A fundamental quantity in linear algebra characterizing a matrix is *rank*.

Definition 3.2.1 (Matrix rank)**.** *Let $M \in \mathbb{R}^{k \times n}$ be a matrix.*

- *The row rank of M is the greatest number of linearly independent rows of M. Equivalently is the dimension of the row space, i.e the vector space spanned by the rows of M.*
- *The column rank of M is the greatest number of linearly independent columns of M. Equivalently is the dimension of the column space, i.e the vector space spanned by the columns of M.*
- *The column rank and the row rank of M are always equal. This quantity is also called rank and is denoted as $rank(M)$.*

3.2.1 Singular value decomposition

A very useful tool in linear algebra is *Singular Value Decomposition (SVD)*, which provides a factorization of a matrix M into three other special matrices. It

has numerous applications, especially in the field of machine learning. Formally, the singular value decomposition of a matrix M is defined as follows.

Theorem 3.2.1 (Singular value decomposition)**.** *Let $M \in \mathbb{R}^{k\times n}$ be a matrix with $rank(M) = r \leq min(k, n)$. The singular value decomposition (SVD) of M is*

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

where

- *$U \in \mathbb{R}^{k\times k}$ is an orthogonal matrix, whose columns are the (normalized) eigenvectors of MM^T (left singular vectors).*
- *$V \in \mathbb{R}^{n\times n}$ is an orthogonal matrix, whose columns are the (normalized) eigenvectors of M^TM (right singular vectors).*
- *$\Sigma \in \mathbb{R}^{k\times n}$ is a (rectangular) diagonal matrix, whose first r diagonal entries are the singular values of M, i.e $\sigma_i = \sqrt{\lambda_i}$, $1 \leq i \leq r$ (λ_i are the non-zero eigenvalues of MM^T), in descending order ($\sigma_1 \geq \sigma_2 \geq \ldots \geq \sigma_r$).*

Also, we can write the SVD of a matrix M as

$$M = U_r\Sigma_r V_r^T,$$

where $U_r \in \mathbb{R}^{k\times r}$ contains the first r columns of U, $V_r \in \mathbb{R}^{n\times r}$ contains the first r columns of V and $\Sigma_r \in \mathbb{R}^{r\times r}$ is a diagonal matrix formulated by inserting only the r non-zero singular values in the diagonal, in descending order. Using this form we can see that the SVD can be expressed as a sum as follows

$$M = \sum_{i=1}^{r} \sigma_i \boldsymbol{u}_k \boldsymbol{v}_k^T, \tag{3.1}$$

where $\boldsymbol{u}_i$, $1 \leq i \leq r$ and $\boldsymbol{v}_i$, $1 \leq i \leq r$ are the first r left and right singular vectors respectively.

We can use the SVD of M to obtain the best l rank (with $l \leq r$) estimate of M, in the Frobenius $\|\cdot\|_F$ and spectral $\|\cdot\|_2$, norm sense. This can be found simply by keeping the first l elements of the SVD expansion (3.1), obtaining

$$\hat{M} = \sum_{i=1}^{l} \sigma_i \boldsymbol{u}_k \boldsymbol{v}_k^T.$$

The theorem that establishes that property of SVD is the *Eckart-Young theorem.*

The time complexity for obtaining the exact SVD of a $k \times n$ matrix is $\mathcal{O}(min\left\{kn^2, k^2n\right\})$. Therefore, computing the SVD is a tractable problem for small and medium size matrices. However, for large data sets or problems where online processing is required, computing the SVD is a difficult task.

Singular value decomposition can reveal several important aspects of low-rank matrices, such as their geometry and the number of parameters that is needed for describing them.[DR16] From a geometric point of view, the set of $k \times n$ matrices with $rank(M) = r$, forms an uncountable union of subspaces in $\mathbb{R}^{k\times n}$. First, notice that every outer product of singular vectors makes a $k \times n$ matrix, i.e $\boldsymbol{u}_i\boldsymbol{v}_i^T \in \mathbb{R}^{k\times n}$. Using expression (3.1) we can see that M is a vector

in the matrix vector space spanned by $\boldsymbol{u}_i\boldsymbol{v}_i^T$, $1 \leq i \leq r \leq kn$ and $\sigma_i, 1 \leq i \leq r$ are the respective expansion coefficients. Taking into account that the set of all $k \times n$ matrices with $rank(M) = r$ accounts for all the different values the $2r$ left and right singular vectors can take (in a continuous domain) and the fact that each different configuration of the singular vectors span a r-dimensional subspace of $\mathbb{R}^{k\times n}$, we conclude that the collection of all $k \times n$ matrices, with $rank(M) = r$, is a union of an uncountable number of r-dimensional subspaces.

Singular value decomposition unveils another intresting aspect. We notice that every element $\sigma_i\boldsymbol{u}_i\boldsymbol{v}_i^T$ is fully specified by $k+n+1$ parameters. As a result, the number of parameters fully characterizing a $k\times n$ matrix, with $rank(M) = r$, are $r(k + n + 1)$. If r is relatively small we have that $r(k + n + 1) << kn$, thus the number of essential parameters of that matrix is much smaller than the number of it's entries. This succinct representation of a low rank matrix is the element that make possible the recovery from a relatively small number of measurements.

3.3 Low-rank matrix recovery

Let $M \in \mathbb{R}^{k\times n}$ be a $k \times n$ matrix. A linear measurement operation on M is modeled as [DR16]

$$\boldsymbol{y} = \mathcal{A}(M) + \boldsymbol{e}, \tag{3.2}$$

where $\boldsymbol{y} \in \mathbb{R}^m$ is the observations/measurements vector, $\boldsymbol{e} \in \mathbb{R}^m$ is the noise vector and $\mathcal{A} : \mathbb{R}^{k\times n} \to \mathbb{R}^m$ is a linear measurement operator that acts as follows:

$$y_i = \left\langle M, A^{(i)} \right\rangle + e_i = tr\left(A^{(i)^T} M\right) + e_i = \sum_{j=1}^{k}\sum_{l=1}^{n} M_{jl}A_{jl}^{(i)} + e_i,\ 1 \leq i \leq m, \tag{3.3}$$

where the $A^{(i)}, 1 \leq i \leq m$ are set of pre-defined matrices and e_i is the respective component of the noise vector.

In contrast to the sparse vector recovery problem, we are going to restrict our attention to special kind of measurements and therefore to specific kinds of recovery problems. The most notable cases are the following : [DR16]

1. **Matrix completion**

   In the matrix completion scenario the matrices $(A^{(i)})_{1\leq i\leq m}$ are defined as

   $$A_{jl}^{(i)} = \begin{cases} 1 & ,(j,l) = (s,t) \\ 0 & ,(j,l) \neq (s,t) \end{cases},\ 1 \leq i \leq m \tag{3.4}$$

   for $(s_i, t_i) \in \{1, \ldots, k\} \times \{1, \ldots, n\}$. Essentially, we have a matrix M and we observe only a subset of it's entries. The objective is to recover the matrix M, given the fact that M is low rank. In other words, the objective is to complete the missing entries of the matrix.

2. **Low-rank matrix recovery from random observations**

   In this scenario each $A^{(i)}$ is a random matrix. The most common case is the one of Gaussian random matrices, where the entries of each matrix are i.i.d Gaussian random variables with mean equal to 0 and variance $\frac{1}{m}$.

3. **Low-rank matrix recovery from rank-1 measurements**

   In this case the matrices $A^{(i)}$ have rank 1. Characteristic examples of problems that belong this category is *phase retrieval* and *blind deconvolution*

The most important and well-studied scenario is matrix completion. Thus, we are going to restrict our attention to that case only.

3.4 Matrix completion

As we mentioned previously, in a *matrix completion* problem we are given a subset of the entries of a matrix $M \in \mathbb{R}^{k \times n}$ and the knowledge that M is low-rank. The objective is to complete successfully the missing entries of M. There are several reasons why a matrix may have missing entries. First, the size of the matrix may be large enough and as result observing the full matrix can be very expensive. Also, it is possible that the missing entries is an intrinsic characteristic of the problem we are dealing with, such as in the case of recommendation systems, where it is practically impossible for every user to rate every possible item.

In general, it is not always possible to complete a low-rank matrix and we are going to study conditions that establish that the matrix can be completed with high probability. Before we do that we must present a characteristic application of matrix completion in order to motivate the development of this subject.

3.4.1 The Netflix problem

The most common example that accompanies a presentation about matrix completion is the Netflix recommendation system problem. We are going to give a brief overview of this problem since it will proven to be useful to have a concrete example to evaluate the soundness of our arguments. In that problem we consider a set of users and a set of movies, where every user can rate any movie in the movies set. We model that recommendation system using a matrix, where each row corresponds to a different user and each column to a different movie. Each entry A_{ij} of the previous matrix corresponds to the rating given by user i to movie j. Obviously, the rating matrix has missing entries, since in a realistic scenario we are talking about thousands of movies and users and therefore it is impractical for every user to rate every possible movie.

The problem that arises is to find a way to fill the missing entries of the matrix, inferring this way the preference of users towards movies they haven't rated (from which a big portion of them most likely correspond to movies they haven't watched yet). In general this problem is impossible to solve. The assumption that renders this task feasible is the low-rank of the rating matrix. The low-rank of the rating matrix reflects the fact that the users often share preferences, as well as movies belonging to the same category may be rated in the same way by different users. The low rank of the rating matrix is the mathematical notion that captures this empirical observation.

3.4.2 Which matrices can be completed?

One of the first thing we must examine is what kind of matrices can be completed. In principle problems arise in the following cases.

- Sparse matrices

  Sparse matrices create a problem as we need to observe the majority of their entries in order to be able to successfully complete them. As a result, it is impractical to complete sparse matrices. For example, consider the following matrix ($*$ denotes the non-zero entry).

$$\begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & * & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix}$$

  It is obvious that if we do not observe the non-zero entry we cannot correctly complete the matrix, as there is no way to know that this element is non-zero.

- Matrices with sparse singular vectors

  Matrices that possess at least one sparse singular vector are problematic. Consider the following example (taken from [FG16]) of a rank 2 matrix.

$$M = \begin{bmatrix} 2 & 2 & 2 & 2 \\ 2 & 2 & 2 & 2 \\ 2 & 2 & 2 & 2 \\ 2 & 2 & 2 & 2 \\ -3 & 3 & -3 & 3 \end{bmatrix} = \sigma_1 u_1 v_1^T + \sigma_2 u_2 v_2^T =$$

$$= 8 \begin{bmatrix} 0.5 \\ 0.5 \\ 0.5 \\ 0.5 \\ 0 \end{bmatrix} \begin{bmatrix} 0.5 & 0.5 & 0.5 & 0.5 \end{bmatrix} + 6 \begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 1 \end{bmatrix} \begin{bmatrix} -0.5 & 0.5 & -0.5 & 0.5 \end{bmatrix}.$$

  Notice that the elements of the singular vectors $\boldsymbol{u}_1, \boldsymbol{v}_1$ contribute to the formation of almost all the rows (rows 1-4), while the elements of the singular vectors $\boldsymbol{u}_2, \boldsymbol{v}_2$ contribute only to the formation of the 5th row. In other words, only the 5th row of the matrix contains information about the singular vectors $\boldsymbol{u}_2, \boldsymbol{v}_2$. Essentially, that means that we need to observe at least all the entries of the 5th row in order to be able to successfully complete the matrix, since the important information of the low rank matrix is contained in its singular vectors and values.

In conclusion, we want the information provided by the singular vectors to spread to many elements of the matrix. As a result, we want spread (or in other words not sparse/spiky) singular vectors. The measure that quantifies how well a matrix M, or more precisely the column and row space of that matrix (the left singular vectors span the column space and the right singular vectors span the row space), adhere to these guidelines is *coherence*.

We also need to examine which sampling sets give completable matrices. Mainly, we have only one problematic case.

- Matrices with at least one row or column missing

  It is impossible to complete a matrix if we fail to observe at least some row or column. The following example contains a matrix with rank 1 and a missing row (* denote the known entries and ? the unknown).

$$\begin{bmatrix} * & * & * & * \\ ? & ? & ? & ? \\ * & * & * & * \\ * & * & * & * \end{bmatrix} = \sigma_1 \begin{bmatrix} * \\ ? \\ * \\ * \end{bmatrix} \begin{bmatrix} * & * & * & * \end{bmatrix}$$

It is apparent that there is no way to infer the missing value in the left singular vector, since this value contributes only to the formation of the 2nd row, whose entries are unknown to us. As a result, we need to observe at least one entry from every column and every row.

We can establish with high probability that we are going to obtain at least one element in every row and column if we apply the uniform sampling model on the entries of M and the number of known entries satisfies some lower bound. Note that in the uniform sampling model every possible subset with m elements of the set of $k \cdot n$ entries of M is picked with equal probability. From the well-known coupon collector's problem we can deduce that we need at least $klogk$ (assuming that $k \geq n$, else we need $nlogn$) known entries in order to ensure with high probability that we have at least one entry in every row and every column.

3.4.3 Coherence

A measure that quantifies the spread of the elements of the singular vectors of a matrix $M \in \mathbb{R}^{k \times n}$ is *coherence*.

Definition 3.4.1 (Coherence)**.** *[CR09] Let U be a subspace of $\mathbb{R}^n$ with $dim U = r$. Also, let P_U denote the orthogonal projection onto U. Then, we define the coherence of U, with respect to the standard basis $(\boldsymbol{e}_i)_{i=1}^n$ as*

$$\mu(U) = \frac{n}{r} \max_{1 \leq i \leq n} \|P_U \boldsymbol{e}_i\|_2^2. \tag{3.5}$$

Notice that $P_U = UU^T$, where U is the matrix formulated by inserting the the basis vectors of U as columns. Thus,

$$\begin{aligned} \|P_U \boldsymbol{e}_i\|_2 = (P_U \boldsymbol{e}_i)^T (P_U \boldsymbol{e}_i) = \boldsymbol{e}_i^T P_U^T P_U \boldsymbol{e}_i = \boldsymbol{e}_i^T UU^T UU^T \boldsymbol{e}_i = \quad \text{(U orthogonal)} \\ = \boldsymbol{e}_i^T UU^T \boldsymbol{e}_i = (U^T \boldsymbol{e}_i)^T (U^T \boldsymbol{e}_i) = \|U^T \boldsymbol{e}_i\|_2^2 \end{aligned}$$

Therefore, we can rewrite coherence as

$$\mu(U) = \frac{n}{r} \max_{1 \leq i \leq n} \|U^T \boldsymbol{e}_i\|^2, \tag{3.6}$$

where U in the right-hand side denotes the subspace matrix, i.e the matrix whose columns are the basis vectors of the aforementioned subspace.

For the values of $\mu(U)$ it holds that

$$1 \leq \mu(U) \leq \frac{n}{r}.$$

The smallest value is attained in cases where all the vectors that span U have entries with magnitude $\frac{1}{\sqrt{n}}$, i.e $u_i = \left(\frac{1}{\sqrt{n}}, \ldots, \frac{1}{\sqrt{n}}\right) \in \mathbb{R}^n$, $1 \leq i \leq r$. The largest value of $\mu(U)$ is attained when there exists $i \in \{1, \ldots, n\}$, such that $\boldsymbol{e}_i$ lies in the span of U.

In the context of matrix completion we are interested in $\mu(U)$ and $\mu(V)$, where U and V are the column and row space respectively of M or equivalently the matrices $U \in \mathbb{R}^{k \times r}$ and $V \in \mathbb{R}^{n \times r}$ of the left and right singular vectors respectively, as provided by the SVD. We want the value of coherence to be as small a possible. This implies that the correlation between the singular vectors with the ("spiky") standard basis vectors is small and as a result the singular vectors are dense and not sparse (i.e "spiky"). Only one sparse singular vector is sufficient to force coherence to attain its maximum value.

3.4.4 The optimization tasks of matrix completion

We proceed to the formulation of the matrix completion problem and its solution. The general setting of the matrix completion problem is the following:

Let $M \in \mathbb{R}^{k \times n}$ be a $k \times n$ matrix with $rank(M) = r$ (approximately) small. We observe m entries of the matrix M picked uniformly at random, while the rest remain unknown. We denote the positions of the known entries as Ω.

The recovery of M is achieved using optimization tasks, as in the case of the sparse vector recovery problem. We are going to provide two different optimization tasks and we will evaluate their capability in accomplishing our objective, i.e successful matrix recovery.

Rank minimization

The optimization task that captures the notion of finding the matrix with the lowest rank is *rank minimization*. The formal definition follows.

Definition 3.4.2 (Rank minimization)**.**

$$\begin{aligned} &\underset{X \in \mathbb{R}^{k \times n}}{\textit{minimize}} && rank(X) \\ &\textit{subject to} && X_{ij} = M_{ij},\ (i,j) \in \Omega. \end{aligned}$$

Essentially, this task searches in the space of all possible $k \times n$ matrices that have the same values as M in the specified positions and picks the solution with the lowest rank. Rank minimization is an NP-Hard problem in the general case, where we consider all possible matrices $M \in \mathbb{R}^{k \times n}$ and all possible sampling sets $\Omega \subseteq \{1, \ldots, k\} \times \{1, \ldots, n\}$ as input. Notice that this task corresponds to the l_0 minimization task used in the sparse vector recovery problem. We can easily notice the correspondence if we see that $rank(X) = \|\sigma(X)\|_0$, where $\sigma(X)$ is the vector of the singular values of matrix X. Both problems are intractable and although they capture the essence of the problem, we reject them and search for tractable alternatives.

Nuclear norm minimization

The search for a tractable alternative leads to an optimization task called *nuclear norm minimization*.

Definition 3.4.3 (Nuclear norm minimization)**.**

$$\begin{aligned} \underset{X\in\mathbb{R}^{k\times n}}{\textit{minimize}} \quad & \|X\|_* \\ \textit{subject to} \quad & X_{ij} = M_{ij},\ (i,j) \subseteq \Omega. \end{aligned}$$

This optimization task can be reduced as a semidefinite optimization problem, hence we can find efficient, polynomial-time, algorithms that can solve it. Notice that this optimization task corresponds to the l_1 minimization task we studied in the previous chapter, since for the nuclear norm we know that $\|X\|_* = \sum_{i=1}^{r} |\sigma_i|$ (remember that σ_i are the singular values of X).

The first provable recovery guarantee using nuclear norm minimization for matrix completion was given in [CR09]. We are going to provide a newer result due to Recht [Rec11], which improves the previous result by providing a better bound for the number of measurements.

Theorem 3.4.1 (Recovery guarantee with nuclear norm minimization)**.** *[Rec11] Let $M \in \mathbb{R}^{k\times n}$ be a matrix of rank r with singular value decomposition $M = U\Sigma V^T$. Without loss of generality, we impose the conventions $m < n$, $\Sigma \in \mathbb{R}^{r\times r}$, $U \in \mathbb{R}^{m\times r}$, $V \in \mathbb{R}^{n\times r}$. Also, we make the following assumptions*

- *The row and column spaces have coherences bounded above by some positive μ_0, i.e $\mu_0 = \max\{\mu(U), \mu(V)\}$.*
- *The matrix UV^T has a maximum entry bounded in absolute value by $\mu_1\sqrt{\frac{r}{mn}}$, where $\mu_1 > 0$.*

  Suppose that we sample uniformly at random m entries from M. Then if

  $$m \geq 32 \max\left\{\mu_1^2, \mu_0\right\} r(k+n)\beta \log^2(2n),$$

  for some $\beta > 1$, the solution of problem (3.4.3) is equal to M with probability at least $1 - 6log(n)(k+n)6^{2-2\beta} - n^{2-2\beta^{1/2}}$.

There are several algorithms that solve the nuclear norm minimization task, as well as algorithms that solve the matrix completion problem using different approaches. Algorithms for matrix completion is the topic of chapter 6, so we will postpone the analysis of some algorithms and the necessary discussion for that part of the thesis.

3.5 Applications

Low-rank matrix recovery has many applications in a wide range of fields in science and engineering. For instance, in literature we encounter the following examples. [DR16]

- **Quantum state tomography**: In quantum mechanics we describe the state of a quantum system using a matrix that is called density matrix. Suppose that we have a particle whose state is described using a 2-dimensional state vector, i.e a qubit. Then the dimension of the state space of a system comprised by n qubits is 2^n and as a result the dimension of the density matrix is $2^n \cdot 2^n = 4^n$. We notice that the dimensions of the density matrix scale exponentially, making the task of finding all the entries of the density matrix and thus the state of the quantum system a difficult task. It is possible when the density matrix is low-rank, i.e the system consists of a small ensemble of pure states, to fill the missing entries of the matrix using matrix completion algorithms. [Gro11]

- **Recommendation systems**: As we mentioned previously, recommendations system is one of the most characteristic examples in the field of matrix completion. Organizing the responses of individuals to a set of items into a matrix, we end up with an incomplete matrix, as it is very difficult to enforce every individual to evaluate every single product. Exploiting the fact that different individuals may share the same preferences, which translates to the low-rank of the corresponding matrix, we apply matrix completion algorithms to fill the missing entries of the ratings matrix.

- **Distance matrices**: In many problems it is useful to have the matrix of pairwise distances between the components of the system we are studying. A characteristic example is a wireless sensor network, which consists of a set of sensors scattered in an area. The respective distance matrix is low rank and may have missing entries, rendering the usage of matrix completion algorithms necessary.

There are many other problems and fields in machine learning and signal processing that contain incomplete low-rank matrices, such as principal component analysis(PCA), natural language processing, multi-task learning, etc.

Part II

Mathematical aspects

Chapter 4

Tools from probability theory

4.1 Probabilistic preliminaries

The purpose of this chapter is to provide all the necessary tools of probability theory that are vital for the next chapter. The tools we develop cover a wide range, starting from elementary definitions and propositions that are part of a typical undergraduate probability course and reaching more advanced topics usually treated within the context of high dimensional probability theory. In this section we are going to provide the most elementary definitions that are necessary for the development of this thesis. This chapter is mainly based on [FR13], essentially comprising a partial presentation of chapters 7 and 8.

The setting in which probability theory takes place is a probability space $(\Omega, \Sigma, \mathbb{P})$, where Ω is the sample space (the set of all possible outcomes/results of a random experiment), Σ is a σ-algebra on the sample space Ω (the set of all possible events) and $\mathbb{P}$ a probability measure on (Ω, Σ) (a function that maps every event to a real number).

Given a probability space $(\Omega, \Sigma, \mathbb{P})$ and $1 \leq p < \infty$, we define the $L_p(\Omega, \Sigma, \mathbb{P})$-space of random variables as the set of random variables on $(\Omega, \Sigma, \mathbb{P})$ with finite L_p norm, i.e

$$L_p(\Omega, \Sigma, \mathbb{P}) = \left\{ X : \Omega \to \mathbb{R} \ : \ \|X\|_p = \left(\mathbb{E}\left[|X|^p\right]\right)^{1/p} < \infty \right\}.$$

Note that for $1 \leq p < \infty$ the L_p space is a Banach space. In the special case where $p = 2$ the L_2 space is a Hilbert space, with the inner product defined as $\langle X, Y \rangle = \mathbb{E}[XY]$.

From the definition of the norm we can immediately obtain the triangle inequality for L_p random variables.

Proposition 4.1.1 (Triangle inequality)**.** *For every $X, Y \in L_p(\Omega, \Sigma, \mathbb{P})$ it holds that*

$$\left(\mathbb{E}\left[|X+Y|^p\right]\right)^{1/p} \leq \left(\mathbb{E}\left[|X|^p\right]\right)^{1/p} + \left(\mathbb{E}\left[|Y|^p\right]\right)^{1/p}$$

for some $1 \leq p < \infty$.

A very useful inequality is *Holder's inequality.*

Proposition 4.1.2 (Holder's inequality). *For every $X \in L_p(\Omega, \Sigma, \mathbb{P})$ and $Y \in L_q(\Omega, \Sigma, \mathbb{P})$ with $p, q \geq 1$, such that $\frac{1}{p} + \frac{1}{q} = 1$ it holds that*

$$|\mathbb{E}[XY]| \leq (\mathbb{E}[|X|^p])^{1/p} \cdot (\mathbb{E}[|Y|^q])^{1/q}.$$

A special case of Holder's inequality, that we can easily obtain by setting $p = q = 2$, is the famous *Cauchy-Schwarz inequality.*

Proposition 4.1.3 (Cauchy-Schwarz inequality). *For every $X, Y \in L_2(\Omega, \Sigma, \mathbb{P})$ it holds that*

$$|\mathbb{E}[XY]| \leq \left(\mathbb{E}\left[X^2\right] \cdot \mathbb{E}\left[Y^2\right]\right)^{1/2}.$$

Another important inequality is *Jensen's inequality.*

Proposition 4.1.4 (Jensen's inequality). *Let $\boldsymbol{X} \in \mathbb{R}^n$ be a random vector and $f : \mathbb{R}^n \to \mathbb{R}$ be a convex function. Then,*

$$f(\mathbb{E}[\boldsymbol{X}]) \leq \mathbb{E}[f(\boldsymbol{X})].$$

One useful function that we are going to use in order to facilitate some proofs is the *characteristic function* of a random variable X.

Definition 4.1.1 (Characteristic function). *The characteristic function of a random variable X on an event A is defined as*

$$\mathbb{I}_{\{X \in A\}}(x) = \begin{cases} 1 & x \in A \\ 0 & x \notin A \end{cases}.$$

Two very useful functions in probability theory are the *moment-generating function* and the *cummulant-generating function.* Essentially the moment-generating function provides an alternative representation of the probability density function of a random variable.

Definition 4.1.2 (Moment-generating function). *The moment-generating function of a (real-valued) random variable X is a function M_X such that*

$$M_X(t) = \mathbb{E}\left[e^{tX}\right], t \in \mathbb{R},$$

whenever this expectation exists.

Definition 4.1.3 (Cummulant-generating function). *The cummulant-generating function of a random variable X is a function C_X such that*

$$C_X(t) = \ln\left(\mathbb{E}\left[e^{tX}\right]\right), t \in \mathbb{R}$$

Lebesgue's dominated convergence theorem is given without a proof.

Theorem 4.1.1 (Lebesgue's dominated convergence theorem). *Let $\{X_n\}_{n \in \mathbb{N}}$ be a sequence of random variables, such that $\lim_{n \to \infty} X_n(\omega) = X(\omega)$, for almost all $\omega \in \Omega$. Also, let Y be a random variable, with $\mathbb{E}[|Y|] < \infty$. If $|X_n| \leq |Y|$, $\forall n \in \mathbb{N}$, almost surely, then*

$$\lim_{n \to \infty} \mathbb{E}[X_n] = \mathbb{E}[X].$$

A generalization of the notion of the random variable is that of a *random vector*, which is a finite collection of random variables.

Definition 4.1.4 (Random vector)**.** *A collection of n random variables,* $\boldsymbol{X} = [X_1, X_2, \ldots, X_n] \in \mathbb{R}^n$*, defined on a common probability space* $(\Omega, \Sigma, \mathbb{P})$ *is called a random vector.*

The type of random vector that we are going to utilize the most is the *standard Gaussian random vector*.

Definition 4.1.5 (Standard Gaussian random vector)**.** *A random vector* $\boldsymbol{g} = [g_1, g_2, \cdots, g_n] \in \mathbb{R}^n$*, such that its components* $g_i, 1 \leq i \leq n$ *are independent standard Gaussian random variables, i.e* $g_i \backsim \mathcal{N}(0, 1)$*, is called a standard Gaussian random vector.*

More generally, about *Gaussian random vectors* we have the following definition.

Definition 4.1.6 (Gaussian random vector)**.** *Let* $\boldsymbol{X} \in \mathbb{R}^n$ *be a random vector defined as* $\boldsymbol{X} = A\boldsymbol{g} + \boldsymbol{\mu}$*, where* $A \in \mathbb{R}^{n \times m}$ *a matrix,* $\boldsymbol{g} \in \mathbb{R}^m$ *a standard Gaussian random vector and* $\boldsymbol{\mu}$ *the expectation of* $\boldsymbol{X}$*. Then,* $\boldsymbol{X} \in \mathbb{R}^n$ *is called a Gaussian random vector.*

The notion of *isotropicity* is roughly a high dimensional generalization of the notion of unit variance in random variables. A characteristic property of isotropic random vectors is provided below.

Definition 4.1.7 (Isotropic random vector)**.** *A random vector* $\boldsymbol{Y} \in \mathbb{R}^n$ *is isotropic if*

$$\mathbb{E}\left[|\langle \boldsymbol{X}, \boldsymbol{x} \rangle|^2\right] = \|\boldsymbol{x}\|_2^2, \, \forall \boldsymbol{x} \in \mathbb{R}^n.$$

A further generalization of a random vector is that of a *random matrix*. In the next chapter part of the results are going to refer to Gaussian random matrices and therefore we provide the formal definition.

Definition 4.1.8 (Gaussian random matrix)**.** *A Gaussian random matrix is a matrix* $A \in \mathbb{R}^{m \times n}$ *whose entries are independent standard Gaussian random variables, i.e* $A_{ij} \backsim \mathcal{N}(0, 1)$*.*

Finally, we give the definition of a *stochastic process* or a *random process*.

Definition 4.1.9 (Stochastic process)**.** *A stochastic process is a collection of random variables, on the same probability space* $(\Omega, \Sigma, \mathbb{P})$*, indexed by some set* T*, i.e* $\{X_t\}_{t \in T}$*.*

4.2 Basic results in probability theory

In this section several fundamental results in probability theory are going to be presented. We provide the majority of them without proofs. The interested reader can refer to [FR13] for more details.

For the expectation of the absolute moments of a random variable we have the following result.

Proposition 4.2.1 (Expectation of absolute moments)**.** *Let X be a random variable. For $p > 0$ it holds that*

$$\mathbb{E}\left[|X|^p\right] = p \int_0^\infty \mathbb{P}\left(|X| \geq t\right) t^{p-1} dt.$$

One of the most famous inequalities in probability theory is *Markov's inequality.*

Theorem 4.2.1 (Markov's inequality)**.** *Let X be a random variable. Then the following inequality holds*

$$\mathbb{P}\left(|X| \geq t\right) \leq \frac{\mathbb{E}\left[|X|\right]}{t}, \forall t > 0.$$

A very important theorem in probability theory is the *central limit theorem*, which underlines the value of the Gaussian distribution.

Theorem 4.2.2 (Central limit theorem)**.** *Let $(X_i)_{i\in\mathbb{N}}$ be a sequence of independent, identically distributed random variables, with $\mathbb{E}\left[X_i\right] = \mu$ and $Var(X_i) = \sigma^2$. Also, consider the following sequence of random variables*

$$Z_n = \frac{\sum_{i=1}^{n} \left(X_i - \mu\right)}{\sigma\sqrt{n}}.$$

Then, the sequence of random variables $(Z_i)_{i\in\mathbb{N}}$ converges in distribution to a standard Gaussian random variable for all bounded continuous function, i.e

$$\lim_{n\to\infty} \mathbb{E}\left[f(Z_n)\right] = \mathbb{E}\left[f(g)\right],$$

where g is a standard Gaussian random variable.

It is useful to have at our disposal the following formula for a standard Gaussian random variable.

Lemma 4.2.1. *Let g be a standard Gaussian random variable. Then, for $t \in \mathbb{R}$ and $c < \frac{1}{2}$ we have that*

$$\mathbb{E}\left[exp(cg^2 + tg)\right] = \frac{1}{\sqrt{1-2c}} exp\left(\frac{t^2}{2(1-2c)}\right). \tag{4.1}$$

The following proposition shows a connection between the moments and the tails of a random variable.

Proposition 4.2.2 (Moments and tails for random variables)**.** *Let X be a random variable that satisfies*

$$\mathbb{P}\left(|X| \geq e^{1/\gamma}\alpha u\right) \leq \beta e^{-u^\gamma/\gamma}, \forall u > 0, \text{ for some } \gamma > 0. \tag{4.2}$$

Then, for $p > 0$ it holds that

$$\mathbb{E}\left[|X|^p\right] \leq \beta a^p (e\gamma)^{p/\gamma} \Gamma\left(\frac{p}{\gamma} + 1\right). \tag{4.3}$$

Also, it holds that

$$\mathbb{E}\left[|X|^p\right]^{1/p} \leq C_1 \alpha C_{2,\gamma}^{1/p} \beta^{1/p} p^{1/\gamma}, \, \forall \, p \geq 1, \tag{4.4}$$

where $C_1 = e^{1/(2e)}$ *and* $C_{2,\gamma} = \sqrt{\dfrac{2\pi}{\gamma}} e^{\gamma/12}$.

An important theorem in probability theory is *Cramer's theorem.*

Theorem 4.2.3 (Cramer's theorem)**.** *Let* $X_1, \ldots, X_n$ *be a finite collection of independent random variables. Then, we have that,*

$$\mathbb{P}\left(\sum_{i=1}^{n} X_i \geq t\right) \leq \exp\left(\inf_{\theta>0}\left\{-\theta t + \sum_{i=1}^{n} C_{X_i}(\theta)\right\}\right), \, \forall t > 0.$$

Proof. For $\theta > 0$, we have that

$$\begin{aligned}
\mathbb{P}\left(\sum_{i=1}^{n} X_i \geq t\right) &= \mathbb{P}\left(\exp\left(\theta \sum_{i=1}^{n} X_i\right) \geq \exp(\theta t)\right) \leq \frac{\mathbb{E}\left[\exp\left(\theta \sum\limits_{i=1}^{n} X_i\right)\right]}{\exp(\theta t)} = && \text{(Markov's inequality)} \\
&= \frac{\mathbb{E}\left[\prod\limits_{i=1}^{n} \exp(\theta X_i)\right]}{\exp(\theta t)} = \frac{\prod\limits_{i=1}^{n} \mathbb{E}\left[\exp(\theta X_i)\right]}{\exp(\theta t)} = && \text{(Independence of } X_i\text{)} \\
&= \frac{\prod\limits_{i=1}^{n} \exp\left(\ln\left(\mathbb{E}\left[\exp(\theta X_i)\right]\right)\right)}{\exp(\theta t)} = \frac{\prod\limits_{i=1}^{n} \exp\left(C_{X_i}(\theta)\right)}{\exp(\theta t)} = \\
&= \exp\left(-\theta t + \sum_{i=1}^{n} C_{X_i}(\theta)\right).
\end{aligned}$$

The previous result holds for an arbitrary $\theta > 0$, so we can conclude that

$$\mathbb{P}\left(\sum_{i=1}^{n} X_i \geq t\right) \leq \inf_{\theta>0}\left\{\exp\left(-\theta t + \sum_{i=1}^{n} C_{X_i}(\theta)\right)\right\} = \exp\left(\inf_{\theta>0}\left\{-\theta t + \sum_{i=1}^{n} C_{X_i}(\theta)\right\}\right).$$

□

4.3 Subgaussian and subexponential random variables

A *subgaussian distribution* is a class of probability distributions, whose tails decay at least as fast as the tails of the Gaussian distribution.

Definition 4.3.1 (Subgaussian random variables)**.** *A random variable* X *is called subgaussian if* $\exists \, k, r > 0$ *such that*

$$\mathbb{P}\left(|X| \geq t\right) \leq k e^{-r t^2}, \, \forall \, t > 0. \tag{4.5}$$

We should mention that Bernoulli, Gaussian and bounded distributions are all special cases of subgaussian distributions. There are several ways to describe subgaussian random variables. One of those ways is contained in the following proposition.

Proposition 4.3.1 (Property of subgaussian random variables)**.** *Let X be subgaussian random variable. Then, there exists $c_1 > 0, c_2 \geq 1$ such that*

$$\mathbb{E}\left[e^{c_1 X^2}\right] \leq c_2.$$

Proof. First of all, we see from definition 4.3.1 that equation (4.2) of proposition 4.2.2 is satisfied for a subgaussian random variable with $\alpha = (2er)^{-1/2}$, $\gamma = 2$, $\beta = k$ and $u = (2r)^{1/2}t$. As a result, we can easily obtain the following estimate for $p = 2n$, using property (A.3) of Gamma functions,

$$\mathbb{E}\left[X^{2n}\right] \leq k(2er)^{-n}(2e)^n \Gamma\left(n+1\right) = kr^{-n}n!.$$

Using the Taylor expansion of the exponential function yields

$$\mathbb{E}\left[e^{c_1 X^2}\right] = 1 + \sum_{n=1}^{\infty} \frac{c_1^n \mathbb{E}\left[X^{2n}\right]}{n!} \leq 1 + k\sum_{n=1}^{\infty} \frac{c_1^n r^{-n} n!}{n!} = 1 + \frac{kc_1 r^{-1}}{1 - c_1 r^{-1}},$$

if we choose c_1 such that $\dfrac{c_1}{r} < 1$. Notice that $c_2 = 1 + \dfrac{kc_1 r}{1 - c_1 r^{-1}} \geq 1$, since $\dfrac{c_1}{r} < 1$ and $k > 0$. □

Also, the following moment estimate for subgaussian random variables will proven to be useful.

Lemma 4.3.1 (Moment estimate for subgaussian random variables)**.** *Let X be a subgaussian random variable. Then,*

$$\left(\mathbb{E}\left[|X|^p\right]\right)^{1/p} \leq Cr^{-1/2}k^{1/p}p^{1/2}, \forall\, p \geq 1,$$

for $C = exp\left(\dfrac{1}{2e} + \dfrac{1}{6}\right)\sqrt{\dfrac{\pi}{2e}}$.

Proof. Using the definition of a subgaussian random variable and proposition 4.2.2 we notice that $\alpha = \dfrac{1}{\sqrt{2er}}$, $\gamma = 2$ and $\beta = k$ (α, γ and β are the parameters introduced in proposition 4.2.2). Hence, substituting these values in equation (4.4) yields

$$\begin{aligned}
\mathbb{E}\left[|X|^p\right]^{1/p} &\leq C_1(2er)^{-1/2}C_{2,2}^{1/p}k^{1/p}p^{1/2} = \\
&= e^{1/(2e)}(2e)^{-1/2}r^{-1/2}(\sqrt{\pi}e^{1/6})^{1/p}k^{1/p}p^{1/2} \leq \\
&\leq e^{1/(2e)}(2e)^{-1/2}\sqrt{\pi}e^{1/6}r^{-1/2}k^{1/p}p^{1/2} \qquad (\sqrt{\pi}e^{1/6} > 1,\, p \geq 1) \\
&= Cr^{-1/2}k^{1/p}p^{1/2},
\end{aligned}$$

where $C = exp\left(\dfrac{1}{2e} + \dfrac{1}{6}\right)\sqrt{\dfrac{\pi}{2e}}$. □

We need an estimate for the moment-generating function of a subgaussian random variable.

Proposition 4.3.2 (Moment-generating function of a subgaussian random variable)**.** *Let X be a subgaussian random variable, with $\mathbb{E}[X] = 0$. Then, we have that*

$$\mathbb{E}\left[e^{tX}\right] \leq e^{ct^2}, \forall\, t \in \mathbb{R}, \tag{4.6}$$

where c is a constant depending only on k and r.

Proof. Let $t \geq 0$. We use the Taylor expansion of the exponential function and the fact that $\mathbb{E}[X] = 0$ to obtain

$$\mathbb{E}\left[e^{tX}\right] = 1 + t\mathbb{E}[X] + \sum_{n=2}^{\infty} \frac{t^n \mathbb{E}[X^n]}{n!} \leq 1 + \sum_{n=2}^{\infty} \frac{t^n \mathbb{E}[|X|^n]}{n!}.$$

First, we consider the case where $0 \leq t \leq t_0$, for some $t_0 \geq 0$ So, we have that

$$\begin{aligned}
\mathbb{E}\left[e^{tX}\right] &\leq 1 + k\sum_{n=2}^{\infty} \frac{C^n t^n r^{-n/2} n^{n/2}}{n!} \leq 1 + \frac{k}{\sqrt{2\pi}} \sum_{n=2}^{\infty} \frac{C^n t^n r^{-n/2} n^{n/2}}{n^n e^{-n}} = && \text{(lemma 4.3.1),(A.5.2)} \\
&= 1 + \frac{k}{\sqrt{2\pi}} \sum_{m=0}^{\infty} \left(Cet_0 r^{-1/2}\right)^{m+2} (m+2)^{-(m+2)/2} \leq && (m = n-2), (t \leq t_0) \\
&\leq 1 + \frac{C^2 t^2 e^2 k}{\sqrt{2\pi} r} \sum_{n=0}^{\infty} \left(Cet_0 r^{-1/2}\right)^{n} = \\
&= 1 + \frac{C^2 t^2 e^2 k}{\sqrt{2\pi} r} \frac{1}{1 - Cet_0 r^{-1/2}} = && \text{(provided } Cet_0 r^{-1/2} < 1) \\
&= 1 + ct^2 \leq e^{ct^2}. && (c = Cet_0 r^{-1/2})
\end{aligned}$$

Next, we move to the case where $t > t_0$. Notice that

$$tX - c't^2 = -\left(\sqrt{c'}t - \frac{X}{2\sqrt{c'}}\right)^2 + \frac{X^2}{4c'} \leq \frac{X^2}{4c'}.$$

Therefore, we can obtain

$$\mathbb{E}\left[e^{tX - c't^2}\right] \leq \mathbb{E}\left[exp\left(\frac{X^2}{4c'}\right)\right].$$

We set $c' = \dfrac{1}{4c_1}$, where $c_1 > 0$ is the constant of proposition 4.3.1. Then, using the same proposition we can conclude that

$$\mathbb{E}\left[e^{tX - c't^2}\right] \leq \mathbb{E}\left[e^{c_1 X^2}\right] \leq c_2.$$

Let define θ as $\theta = ln(c_2) t_0^{-2}$, then

$$\begin{aligned}
\mathbb{E}\left[e^{tX}\right] &\leq c_2 e^{c't^2} = c_2 e^{-\theta t^2} e^{(c'+\theta)t^2} \leq \\
&\leq c_2 e^{-\theta t_0^2} e^{(c'+\theta)t^2} = e^{(c'+\theta)t^2}.
\end{aligned}$$

Thus, if we set $c_0 = max\{c, c' + \theta\}$ we have established that

$$\mathbb{E}\left[e^{tX}\right] \leq e^{ct^2}, \forall\, t \geq 0$$

Finally, if we exchange X with $-X$ we can easily see that the previous inequality also holds for $t < 0$. □

Note that any constant c that satisfies equation (4.6) is called a *subgaussian parameter* of the subgaussian random variable X, however we prefer to use as c the smallest possible value. From now on, when we refer to a subgaussian random variable with parameter c, we mean the parameter c that is involved in equation (4.6).

The next theorem essentially states that the distribution of a sum of independent subgaussian random variables remains subgaussian.

Proposition 4.3.3 (Sum of subgaussian random variables)**.** *Let $X_1, \ldots, X_n$ be a sequence of independent subgaussian random variables, with $\mathbb{E}[X_i] = 0$, $1 \leq i \leq n$ and subgaussian parameter c. For $\boldsymbol{a} \in \mathbb{R}^n$ it holds that the random variable*

$$Y = \sum_{i=1}^{n} a_i X_i$$

is subgaussian with parameter $c\|\boldsymbol{a}\|_2^2$

Proof. We know that $X_1, \ldots, X_n$ are subgaussian random variables with $\mathbb{E}[X_i] = 0$, $1 \leq i \leq n$ and subgaussian parameter c, so using proposition 4.3.2 we can obtain

$$\mathbb{E}\left[e^{tX_i}\right] \leq e^{ct^2}, \forall t \in \mathbb{R}, 1 \leq i \leq n. \tag{4.7}$$

For the random variable Y we have,

$$\begin{aligned}
\mathbb{E}\left[e^{tY}\right] &= \mathbb{E}\left[exp\left(t\sum_{i=1}^{n} a_i X_i\right)\right] = \mathbb{E}\left[\prod_{i=1}^{n} e^{ta_i X_i}\right] = \\
&= \prod_{i=1}^{n} \mathbb{E}\left[e^{ta_i X_i}\right] \leq \prod_{i=1}^{n} e^{ca_i^2 t^2} = \qquad \text{(Independence), (4.7)} \\
&= e^{c\|\boldsymbol{a}\|_2^2 t^2}.
\end{aligned}$$

As a result, the random variable Y is subgaussian, with parameter $c\|\boldsymbol{a}\|_2^2$. □

Next we provide the definition of a *subgaussian random vector*. Essentially, a random vector $\boldsymbol{X} \in \mathbb{R}^n$ follows a subgaussian distribution if all of it's one dimensional marginals $\langle \boldsymbol{X}, \boldsymbol{x} \rangle$ also follow a subgaussian distribution.

Definition 4.3.2 (Subgaussian random vector)**.** *A random vector $\boldsymbol{X} \in \mathbb{R}^n$ is called a subgaussian random vector if, for all $\boldsymbol{x} \in \mathbb{R}^n$ with $\|\boldsymbol{x}\|_2 = 1$, the random variable $\langle \boldsymbol{X}, \boldsymbol{x} \rangle$ is subgaussian, with subgaussian parameter c (independent of $\boldsymbol{x}$).*

Again here, when we refer to a subgaussian random vector with parameter c, we mean the parameter c that is contained in definition 4.3.2.

A large part of the presentation that is going to follow uses subgaussian random matrices. So, we provide a rigorous definition of a subgaussian random matrix.

Definition 4.3.3 (Subgaussian random matrix)**.** *A subgaussian random matrix is a matrix $A \in \mathbb{R}^{m\times n}$, whose entries $A_{i,j}$ are independent subgaussian random variables, with $\mathbb{E}[A_{i,j}] = 0$, $Var[A_{i,j}] = 1$ and the same subgaussian parameters k, r.*

Note that the entries of a subgaussian random vector, as well as the entries of a subgaussian random matrix are not necessarily identically distributed.

A *subexponential distribution* is a class of probability distributions, whose tails decay more slowly than any exponential tail. It is a wide class of distributions accounting for subgaussian distributions and some other distributions whose tails are heavier than Gaussian tails.

Definition 4.3.4 (Subexponential random variable)**.** *A random variable X is called subexponential if $\exists\, k, r > 0$ such that*

$$\mathbb{P}(|X| \geq t) \leq ke^{-rt}, \forall\, t > 0.$$

4.4 Bernstein's inequality

One important topic in probability theory is concentration inequalities. Roughly, a *concentration inequality* is an inequality that quantifies the deviation of a random variable X around its expectation $\mathbb{E}[X]$ (or more generally around some other variable), i.e $\mathbb{P}[|X - \mathbb{E}[X]| \geq t] \leq f(t)$, for some function f of t. One important tool we need for the next chapter that falls under the scope of concentration inequalities is *Bernstein's inequality.*

Theorem 4.4.1 (Bernstein's inequality)**.** *Let $X_1, \ldots, X_n$ be independent random variables, with $\mathbb{E}[X_i] = 0, 1 \leq i \leq n$ and*

$$\mathbb{E}[|X_i|^m] \leq m! c^{m-2}\sigma_i^2/2,\ 1 \leq i \leq n,\ m \geq 2, \tag{4.8}$$

where $c > 0$ and $\sigma_i > 0$, $1 \leq i \leq n$ are constants. Then the following holds,

$$\mathbb{P}\left(\left|\sum_{i=1}^{n} X_i\right| \geq t\right) \leq 2\exp\left(-\frac{t^2}{2(\sigma^2 + ct)}\right), \forall\, t > 0,$$

where $\sigma^2 = \sum_{i=1}^{n} \sigma_i^2$.

Proof. Using the Taylor expansion of the exponential function and taking into account that $\mathbb{E}[\theta X_i] = 0$ we get

$$\mathbb{E}[\exp(\theta X_i)] = 1 + \theta\mathbb{E}[X_i] + \sum_{m=2}^{\infty} \frac{\theta^m \mathbb{E}[X_i^m]}{m!} = 1 + \frac{\theta^2\sigma_i^2}{2} \sum_{m=2}^{\infty} \frac{2\theta^{m-2}\mathbb{E}[X_i^m]}{m!\sigma_i^2},\ 1 \leq i \leq n \tag{4.9}$$

For clarity we replace the summation with the following expression

$$S_i(\theta) = \sum_{m=2}^{\infty} \frac{2\theta^{m-2}\mathbb{E}[X_i^m]}{m!\sigma_i^2},\ 1 \leq i \leq n.$$

Using the well-known inequality $1 + x \leq e^x$ and (4.9) we can obtain

$$\mathbb{E}\left[\exp\left(\theta X_i\right)\right] = 1 + \frac{\theta^2\sigma_i^2}{2}S_i(\theta) \leq \exp\left(\frac{\theta^2\sigma_i^2}{2}S_i(\theta)\right) \leq \exp\left(\frac{\theta^2\sigma_i^2}{2}S(\theta)\right),\ 1 \leq i \leq n, \tag{4.10}$$

where $S(\theta) = \max_{1\leq i\leq n} S_i(\theta)$.
For the cummulant-generating function of X_i we have that

$$C_{X_i}(\theta) = \ln\left(\mathbb{E}\left[\exp\left(\theta X_i\right)\right]\right) \leq \frac{\theta^2\sigma_i^2 S(\theta)}{2},\ 1 \leq i \leq n, \tag{4.11}$$

where we have used inequality (4.10).
So Cramer's theorem (4.2.3) and expression (4.11) yields

$$\begin{aligned}\mathbb{P}\left(\sum_{i=1}^n X_i \geq t\right) &\leq \exp\left(\inf_{\theta>0}\left\{-\theta t + \sum_{i=1}^n C_{X_i}(\theta)\right\}\right) \leq \inf_{\theta>0}\left\{\exp\left(-\theta t + \sum_{i=1}^n \frac{\theta^2\sigma_i^2 S(\theta)}{2}\right)\right\} = \\ &= \inf_{\theta>0}\left\{\exp\left(-\theta t + \frac{\theta^2 S(\theta)}{2}\sum_{i=1}^n \sigma_i^2\right)\right\} = \inf_{\theta>0}\left\{\exp\left(-\theta t + \frac{\theta^2\sigma^2 S(\theta)}{2}\right)\right\} \leq \ , \\ &\leq \inf_{0<c\theta<1}\left\{\exp\left(-\theta t + \frac{\theta^2\sigma^2 S(\theta)}{2}\right)\right\}\end{aligned} \tag{4.12}$$

where c is the constant defined in (4.8).
For $0 < c\theta < 1$ it holds that

$$S_i(\theta) = \sum_{m=2}^{\infty}\frac{2\theta^{m-2}\mathbb{E}[X_i^m]}{m!\sigma_i^2} \leq \sum_{m=2}^{\infty}\frac{2\theta^{m-2}\mathbb{E}[|X_i|^m]}{m!\sigma_i^2} \leq \sum_{m=2}^{\infty}\left(\frac{1}{c\theta}\right)^{m-2} = \frac{1}{1-c\theta},\ 1 \leq i \leq n,$$

where we have used the moment bound (4.8).
As a result,

$$S(\theta) \leq \frac{1}{1-c\theta},\ 0 < c\theta < 1$$

The combination of the previous result with expression (4.12) leads to

$$\mathbb{P}\left(\sum_{i=1}^n X_i \geq t\right) \leq \inf_{0<c\theta<1}\left\{\exp\left(-\theta t + \frac{\theta^2\sigma^2}{2\left(1-c\theta\right)}\right)\right\}$$

Finally, by considering $\theta = \dfrac{t}{\sigma^2 + ct}$, which satisfies $c\theta < 1$, we have that

$$\begin{aligned}\mathbb{P}\left(\sum_{i=1}^n X_i \geq t\right) &\leq \exp\left(-\frac{t}{\sigma^2+ct}\cdot t + \frac{\left(\frac{t}{\sigma^2+ct}\right)^2\sigma^2}{2\left(1 - c\cdot\frac{t}{\sigma^2+ct}\right)}\right) = \\ &= \exp\left(\frac{t^2}{2(\sigma^2+ct)}\right).\end{aligned}$$

Exchanging X_i with $-X_i$ in the final expression yields

$$\mathbb{P}\left(\sum_{i=1}^{n} X_i \leq -t\right) \leq \exp\left(\frac{t^2}{2(\sigma^2 + ct)}\right).$$

Consequently, using the union bound we obtain

$$\mathbb{P}\left(\left|\sum_{i=1}^{n} X_i\right| \geq t\right) \leq 2\exp\left(-\frac{t^2}{2(\sigma^2 + ct)}\right), \forall t > 0.$$

□

We can use Bernstein's inequality for subgaussian random variables in order to prove that the same inequality holds for zero mean subexponential random variables.

Corollary 4.4.1 (Bernstein's inequality for subexponential random variables)**.** *Let $X_1, \ldots, X_n$ be independent subexponential random variables, with $\mathbb{E}[X_i] = 0$, $1 \leq i \leq n$. Then, for every $t > 0$*

$$\mathbb{P}\left(\left|\sum_{i=1}^{n} X_i\right| \geq t\right) \leq 2\exp\left(-\frac{r^2t^2}{4kn + 2rt}\right).$$

Proof. For $n \in \mathbb{N}$, $n \geq 2$ we have,

$$\begin{aligned}
\mathbb{E}[|X_i|^n] &= n\int_0^\infty \mathbb{P}(|X_i| \geq t)t^{n-1}dt \leq kn\int_0^\infty e^{-rt}t^{n-1}dt = && (pro.\ 4.2.1),\ (def.\ 4.3.4)\\
&= knr^{-n}\int_0^\infty e^{-u}u^{n-1}du = knr^{-n}\Gamma(n-1) = && (u = rt \Rightarrow du = rdt)\\
&= knr^{-n}(n-1)! = kn!r^{-n} = n!r^{-(n-2)}\frac{2kr^2}{2} && (def.A.3.1)
\end{aligned}$$

Thus, the above expression satisfies condition (4.8) with $c = r^{-1}$ and $\sigma_i^2 = 2kr^{-2}$. As a result, Bernstein's inequality (theorem 4.4.1) holds for the zero-mean subexponential random variables defined in the beginning. □

4.5 Expectation of norms of Gaussian vectors

In this section we are going to prove some results concerning the expectation of the l_2 norm of a Gaussian random vector. In order to do that we need some intermediate results.

Lemma 4.5.1 (Pdf of sum of independent random variables)**.** *The probability density function of the sum $X + Y$ of two independent random variables X and Y with probability density functions p_X and p_Y respectively is given by*

$$p_{X+Y}(z) = (p_X * p_Y)(z) = \int_{-\infty}^{\infty} p_X(u)p_Y(z-u)du$$

Two necessary tools in order to estimate the expectation of the norm of a Gaussian random vector are Beta and Gamma functions. For the respective definitions, as well as useful properties that we are going to use, refer to subsection A.3.1 of the appendix.

Proposition 4.5.1. *Let* $\boldsymbol{g} = [g_1, g_2, \cdots, g_n] \in \mathbb{R}^n$ *be a standard Gaussian random vector. The random variable*

$$Y = \|\boldsymbol{g}\|_2^2 = \sum_{i=1}^{n} g_i^2$$

follows a $\chi^2(n)$ *distribution. The probability density function of the latter is*

$$p_{\chi^2(n)}(u) = \begin{cases} \dfrac{1}{2^{n/2}\Gamma(n/2)} u^{(n/2)-1} e^{-u/2} & , u > 0 \\ 0 & , u \leq 0 \end{cases} \tag{4.13}$$

Proof. We are going to use induction in the number n of elements of $\boldsymbol{g}$. For $n = 1$ we have

$$\mathbb{P}(g^2 \leq u) = \begin{cases} \mathbb{P}(-\sqrt{u} \leq g \leq \sqrt{u}) & , u \geq 0 \\ 0 & , u < 0 \end{cases} = \begin{cases} F_{\mathcal{N}(0,1)}(\sqrt{u}) - F_{\mathcal{N}(0,1)}(-\sqrt{u}) & u \geq 0 \\ 0 & u < 0 \end{cases}.$$

where $F_{\mathcal{N}(0,1)}$ is the cumulative density function of a standard Gaussian random variable.

The probability density function of the random variable g^2 is

$$p_1(u) = \frac{d}{du}\left(F_{\mathcal{N}(0,1)}(\sqrt{u}) - F_{\mathcal{N}(0,1)}(-\sqrt{u})\right) = \frac{1}{2\sqrt{u}} p_{\mathcal{N}(0,1)}(\sqrt{u}) + \frac{1}{2\sqrt{u}} p_{\mathcal{N}(0,1)}(-\sqrt{u}) =$$
$$= \frac{1}{2\sqrt{u}}\left(\frac{1}{\sqrt{2\pi}} e^{-u/2} + \frac{1}{\sqrt{2\pi}} e^{-u/2}\right) = \frac{1}{2\sqrt{u}} u^{-1/2} e^{-u/2},$$

for $u \geq 0$ and $p_1 = 0$, for $u < 0$.

Note that $\Gamma\left(\frac{1}{2}\right) = \sqrt{\pi}$ (eq. A.4), so expression (4.13) holds for $n = 1$.

Suppose that expression (4.13) holds for some$n > 1$. We are going to show that it holds for $n + 1$. Obviously, for $u \leq 0$ we can deduce that $p_{n+1}(u) = 0$. For $u > 0$ we can use lemma 4.5.1 to obtain

$$p_{n+1}(u) = (p_n * p_1)(u) = \int_{-\infty}^{\infty} p_n(t)p_1(u-t)dt = \qquad (t > 0 \text{ and } u - t > 0)$$

$$= \frac{1}{2^{n/2}2^{1/2}\Gamma(n/2)\Gamma(1/2)} \int_0^u t^{(n/2)-1}e^{-t/2}(u-t)^{-1/2}e^{-(u-t)/2}dt =$$

$$= \frac{1}{2^{(n+1)/2}\Gamma(n/2)\Gamma(1/2)}e^{-u/2} \int_0^u t^{(n/2)-1}(u-t)^{-1/2}dt = \qquad (\text{Set } x = \frac{t}{u})$$

$$= \frac{1}{2^{(n+1)/2}\Gamma(n/2)\Gamma(1/2)}e^{-u/2}u^{(n/2)-(1/2)} \int_0^1 x^{(n/2)-1}(1-x)^{-1/2}dx = \qquad (0 < x < 1)$$

$$= \frac{1}{2^{(n+1)/2}\Gamma(n/2)\Gamma(1/2)}e^{-u/2}u^{(n+1)/2-1}B\left(\frac{n}{2}, \frac{1}{2}\right) = \qquad (\text{def. A.3.2})$$

$$= \frac{1}{2^{(n+1)/2}\Gamma(n/2)\Gamma(1/2)}e^{-u/2}u^{(n+1)/2-1}\frac{\Gamma\left(\frac{n}{2}\right)\Gamma\left(\frac{1}{2}\right)}{\Gamma\left(\frac{n+1}{2}\right)} =$$

$$= \frac{1}{2^{(n+1)/2}\Gamma((n+1)/2)}u^{(n+1)/2-1}e^{-u/2}.$$

As a result, the formula (4.13) holds for every $n \in \mathbb{N}$. □

The following theorem contains some bounds for the expectation of the l_2 norm of a Gaussian random vector.

Theorem 4.5.1. *Let $\boldsymbol{g} = [g_1, g_2, \cdots, g_n] \in \mathbb{R}^n$ be a standard Gaussian random vector. Then,*

$$\frac{n}{\sqrt{n+1}} \leq \mathbb{E}\left[\|\boldsymbol{g}\|_2\right] = \sqrt{2}\frac{\Gamma\left(\frac{n+1}{2}\right)}{\Gamma\left(\frac{n}{2}\right)} \leq \sqrt{n}$$

Proof. We know from proposition (4.5.1) that the random variable $\|\boldsymbol{g}\|_2^2$ follows the $\chi^2(n)$ distribution, with probability density function $p_{\chi^2(n)}(u)$, given in equation (4.13). So, we have that

$$\mathbb{E}\left[\|\boldsymbol{g}\|_2\right] = \mathbb{E}\left[\left(\|\boldsymbol{g}\|_2^2\right)^{1/2}\right] = \int_0^\infty u^{1/2}p_{\chi^2(n)}(u)du = \frac{1}{2^{n/2}\Gamma(n/2)}\int_0^\infty u^{1/2}u^{n/2-1}e^{-u/2}du = \quad (t = \frac{u}{2})$$

$$= \frac{1}{2^{n/2}\Gamma(n/2)}\int_0^\infty (2t)^{1/2}(2t)^{n/2-1}e^{-t}2dt = \frac{2^{n/2+1/2}}{2^{n/2}\Gamma(n/2)}\int_0^\infty t^{n/2-1/2}e^{-t}dt =$$

$$= \frac{2^{1/2}}{\Gamma(n/2)}\int_0^\infty t^{((n+1)/2)-1}e^{-t}dt = \sqrt{2}\frac{\Gamma\left(\frac{n+1}{2}\right)}{\Gamma\left(\frac{n}{2}\right)}$$

Since $g_i \sim \mathcal{N}(0,1)$, $1 \leq i \leq n$, it is straightforward that

$$\mathbb{E}\left[\|\boldsymbol{g}\|_2^2\right] = \sum_{i=1}^{n} \mathbb{E}\left[g_i^2\right] = n.$$

Jensen's inequality (definition 4.1.4) provides us the upper bound

$$\begin{aligned}\left(\mathbb{E}\left[\|\boldsymbol{g}\|_2\right]\right)^2 \leq \mathbb{E}\left[\|\boldsymbol{g}\|_2^2\right] &\Rightarrow \mathbb{E}\left[\|\boldsymbol{g}\|_2\right] \leq \sqrt{\mathbb{E}\left[\|\boldsymbol{g}\|_2^2\right]} \\ &\Rightarrow \mathbb{E}\left[\|\boldsymbol{g}\|_2\right] \leq \sqrt{n}.\end{aligned}$$

Let E_n denote the expectation of the l_2 norm of a Gaussian random vector $\boldsymbol{g} \in \mathbb{R}^n$, i.e

$$E_n = \mathbb{E}\left[\|\boldsymbol{g}\|_2\right].$$

Then, using equation (A.2) yields

$$E_{n+1}E_n = \sqrt{2}\frac{\Gamma\left(\frac{n+2}{2}\right)}{\Gamma\left(\frac{n+1}{2}\right)}\sqrt{2}\frac{\Gamma\left(\frac{n+1}{2}\right)}{\Gamma\left(\frac{n}{2}\right)} = 2\frac{\Gamma\left(\frac{n}{2}+1\right)}{\Gamma\left(\frac{n}{2}\right)} = 2\frac{\frac{n}{2}\Gamma\left(\frac{n}{2}\right)}{\Gamma\left(\frac{n}{2}\right)} = n.$$

Finally, the previous expression combined with $E_{n+1} \leq \sqrt{n+1}$ yields the lower bound

$$E_n = \frac{n}{E_{n+1}} \geq \frac{n}{\sqrt{n+1}}.$$

□

4.6 Gaussian width

An important geometric quantity in high dimensional geometry, characterizing a subset $T \subseteq \mathbb{R}^n$, is *Gaussian width*. The definition of Gaussian width follows.

Definition 4.6.1 (Gaussian width)**.** *Let $T \subseteq \mathbb{R}^n$ be a subset of $\mathbb{R}^n$ and $\boldsymbol{g} \in \mathbb{R}^n$ be a standard Gaussian random vector. Then, the Gaussian width of the set T is defined as*

$$w(T) = \mathbb{E}\left[\sup_{\boldsymbol{x}\in T} \langle \boldsymbol{g}, \boldsymbol{x}\rangle\right].$$

Roughly, Gaussian width provides a measure of the width of some set $T \subseteq \mathbb{R}^n$ averaged over the set of directions defined by a standard Gaussian random vector.

Next, we provide some basic properties of Gaussian width. [Cha+12b]

Proposition 4.6.1 (Properties of Gaussian width)**.** *Let $T \subseteq \mathbb{R}^n$. Then for the Gaussian width $w(T)$ of the set T the following properties hold*

(a) $w(T)$ is finite $\iff$ T is bounded

(b) Invariance under translations

(c) Invariance under unitary transformations

(d) Homogeneity, i.e $w(cT) = cw(T)$, $\forall\, c > 0$

(e) Monotonicity, i.e if $T_1 \subseteq T_2$ *then* $w(T_1) \leq w(T_2)$

(f) The Gaussian width of a set is equal to the Gaussian width of it's convex hull, i.e $w(T) = w(conv(T))$

4.7 Gordon's Lemma

The purpose of this section is to prove *Gordon's lemma*. This lemma can be classified to the part of high-dimensional probability theory that aims at bounding the expectation of the supremum of a stochastic process, i.e $\mathbb{E}\left[\sup_{\boldsymbol{x}\in T} X_t\right]$. Note that we use the notion of lattice supremum, i.e

$$\mathbb{E}\left[\sup_{\boldsymbol{x}\in T} X_t\right] = \sup_{\substack{T'\subseteq T \\ T' \text{ finite}}} \left\{\mathbb{E}\left[\sup_{\boldsymbol{x}\in T'} X_t\right]\right\}.$$

Using the lattice supremum helps avoiding measurability issues that may arise in the calculation of the supremum of an uncountable set of random variables.

The underlying notion behind Gordon's lemma is that the relation between the expectations of functions of two families of Gaussian random variables are determined by the relation of their respective covariances, since the latter completely characterizes the distribution of a mean-zero Gaussian random vector. In order to prove Gordon's lemma we need several intermediate lemmas and propositions. First, we give the definition of a *function of moderate growth*.

Definition 4.7.1 (Functions of moderate growth)**.** *Let* $F : \mathbb{R}^n \to \mathbb{R}$ *be a function. We say that* F *is of moderate growth if for every* $c > 0$*, it holds that*

$$\lim_{\|\boldsymbol{x}\|_2\to\infty} F(\boldsymbol{x})e^{-c\|\boldsymbol{x}\|_2^2} = 0.$$

The first result that we are going to employ is *Stein's lemma*, also known as *Gaussian integration by parts formula*.

Proposition 4.7.1 (Stein's Lemma)**.** *Let* $F : \mathbb{R}^n \to \mathbb{R}$ *be a differentiable function. Also, suppose that* F *and its first-order partial derivatives are of moderate growth. Then the following statements hold*

1. *Let* g *be a Gaussian random variable, with* $\mathbb{E}[g] = 0$*, and* $n = 1$*. It follows that*
$$\mathbb{E}[gF(g)] = \mathbb{E}\left[g^2\right]\mathbb{E}\left[F'(g)\right].$$

2. *Let* $\boldsymbol{g} = (g_1, \ldots, g_n)$ *be a zero-mean Gaussian random vector and* g_0 *be a zero-mean Gaussian random variable (not necessarily independent of* $\boldsymbol{g}$*). Then it holds that*
$$\mathbb{E}[g_0 F(\boldsymbol{g})] = \sum_{i=1}^{n} \mathbb{E}[g_0 g_i]\,\mathbb{E}\left[\frac{\vartheta F}{\vartheta x_i}(\boldsymbol{g})\right].$$

Proof. 1. For the random variable g we know that $\mathbb{E}[g] = 0$, so $\sigma^2 = \mathbb{E}\left[g^2\right]$. Then using integration by parts we can find that

$$\mathbb{E}[gF(g)] = \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}\sigma} exp\left(-\frac{u^2}{2\sigma^2}\right) \cdot uF(u)du =$$

$$= \frac{1}{\sqrt{2\pi}\sigma} \int_{-\infty}^{\infty} \left(-\sigma^2 exp\left(-\frac{u^2}{2\sigma^2}\right)\right)' F(u)du =$$

$$= -\frac{1}{\sqrt{2\pi}\sigma}\sigma^2 \left[exp\left(-\frac{u^2}{2\sigma^2}\right) F(u)\Big|_{-\infty}^{\infty} - \int_{-\infty}^{\infty} exp\left(-\frac{u^2}{2\sigma^2}\right) F'(u)du\right] =$$

$$= \sigma^2 \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}\sigma} exp\left(-\frac{u^2}{2\sigma^2}\right) F'(u)du = \mathbb{E}\left[g^2\right]\mathbb{E}\left[F'(g)\right].$$

The fact that F is of moderate growth (with $c = \frac{1}{2\sigma^2}$) is what establishes that

$$exp\left(-\frac{u^2}{2\sigma^2}\right) F(u)\Big|_{-\infty}^{\infty} = 0.$$

2. Let $g_i' = g_i - g_0\frac{\mathbb{E}[g_0 g_i]}{\mathbb{E}[g_0^2]}$, $1 \leq i \leq n$ be n random variables.

Also, consider the function $G : \mathbb{R} \to \mathbb{R}$ with

$$G(t) = F\left(g_1' + t\frac{\mathbb{E}[g_0 g_1]}{\mathbb{E}[g_0^2]}, \ldots, g_n' + t\frac{\mathbb{E}[g_0 g_n]}{\mathbb{E}[g_0^2]}\right).$$

Then, if we condition on $\boldsymbol{g}'$ we have that

$$\mathbb{E}[g_0 F(\boldsymbol{g})] = \mathbb{E}[g_0 F(g_1, \ldots, g_n)] = \mathbb{E}\left[g_0 F\left(g_1' + g_0\frac{\mathbb{E}[g_0 g_1]}{\mathbb{E}[g_0^2]}, \ldots, g_n' + g_0\frac{\mathbb{E}[g_0 g_n]}{\mathbb{E}[g_0^2]}\right)\right] =$$

$$= \mathbb{E}[g_0 G(g_0)] = \mathbb{E}\left[g_0^2\right]\mathbb{E}[G'(g_0)] =$$

$$= \mathbb{E}\left[g_0^2\right]\mathbb{E}\left[\frac{dF}{dt}\left(g_1' + t\frac{\mathbb{E}[g_0 g_1]}{\mathbb{E}[g_0^2]}, \ldots, g_n' + t\frac{\mathbb{E}[g_0 g_n]}{\mathbb{E}[g_0^2]}\right)\Big|_{t=g_0}\right] =$$

$$= \mathbb{E}\left[g_0^2\right]\mathbb{E}\left[\sum_{i=1}^{n} \frac{\vartheta F}{\vartheta x_i}\left(g_1' + t\frac{\mathbb{E}[g_0 g_1]}{\mathbb{E}[g_0^2]}, \ldots, g_n' + t\frac{\mathbb{E}[g_0 g_n]}{\mathbb{E}[g_0^2]}\right)\frac{dx_i}{dt}\Big|_{t=g_0}\right] =$$

$$= \mathbb{E}\left[g_0^2\right]\mathbb{E}\left[\sum_{i=1}^{n} \frac{\vartheta F}{\vartheta x_i}\left(g_1' + g_0\frac{\mathbb{E}[g_0 g_1]}{\mathbb{E}[g_0^2]}, \ldots, g_n' + g_0\frac{\mathbb{E}[g_0 g_n]}{\mathbb{E}[g_0^2]}\right)\frac{\mathbb{E}[g_0 g_i]}{\mathbb{E}[g_0^2]}\right] =$$

$$= \mathbb{E}\left[g_0^2\right]\sum_{i=1}^{n}\mathbb{E}\left[\frac{\vartheta F}{\vartheta x_i}(g_1, \ldots, g_n)\frac{\mathbb{E}[g_0 g_i]}{\mathbb{E}[g_0^2]}\right] =$$

$$= \sum_{i=1}^{n}\mathbb{E}[g_0 g_i]\,\mathbb{E}\left[\frac{\vartheta F}{\vartheta x_i}(\boldsymbol{g})\right].$$

□

A characteristic result in integration theory that we are going to use is given in the following proposition.

Proposition 4.7.2. *Let $J \subseteq \mathbb{R}$ be an open interval and $f : J \times \Omega \to \mathbb{R}$ be a function defined on it. Also, let $X : \Omega \to \mathbb{R}$ be a random variable, such that the mapping $t \mapsto f(t, X)$ is almost surely continuously differentiable in J. If for every compact subinterval $I \subset J$,*

$$\mathbb{E}\left[\sup_{t\in I} |f'(t, X)|\right] < \infty \tag{4.14}$$

then the function $g : \mathbb{R} \to \mathbb{R}$, with $g(t) = \mathbb{E}[f(t, X)]$, is continuously differentiable and

$$g'(t) = \mathbb{E}[f'(t, X)].$$

Proof. Let $t \in intJ$ and $I \subset J$ a compact subinterval containing t in its interior. For some $h \in \mathbb{R} \setminus \{0\}$ consider the set $[t, t + h]$ (or $[t - h, t]$ in the case where $h < 0$), such that $[t, t + h] \subseteq I$. Then, the mean value theorem establishes that there exists $\zeta \in [t, t + h]$, such that

$$f'(\zeta, X) = \frac{f(t + h, X) - f(t, X)}{h} = f_h(t, X).$$

It is obvious that

$$|f'(\zeta, X)| \leq \sup_{t\in I} |f'(t, X)| \Rightarrow |f_h(t, X)| \leq \sup_{t\in I} |f'(t, X)|.$$

Also, we know that $\mathbb{E}\left[\sup_{t\in I} |f'(t, X)|\right] < \infty$.

As a result, by Lebesgue's dominated convergence theorem (Theorem (4.1.1))) we get that

$$\lim_{h\to 0} \mathbb{E}[f_h(t, X)] = \mathbb{E}\left[\lim_{h\to 0} f_h(t, X)\right] = \mathbb{E}[f'(t, X)].$$

Using the definition of $g(t)$ we know that

$$\begin{aligned} g'(t) &= \lim_{h\to 0} g_h(t) = \lim_{h\to 0} \frac{g(t + h) - fg(t)}{h} = \\ &= \lim_{h\to 0} \frac{\mathbb{E}[f(t + h, X)] - \mathbb{E}[f(t, X)]}{h} = \lim_{h\to 0} \mathbb{E}[f_h(t, X)]. \end{aligned}$$

Combining the previous two expressions yields,

$$g'(t) = \mathbb{E}[f'(t, X)].$$

□

The next proposition contains a vital tool for the proof of Gordon's lemma.

Proposition 4.7.3. *Let $F : \mathbb{R}^n \to \mathbb{R}$ be a differentiable function, with F and all its first order partial derivatives to be of moderate growth. Also, let $\boldsymbol{X} =$*

$(X_1, \dots, X_n)$ and $\boldsymbol{Y} = (Y_1, \dots, Y_n)$ be two independent Gaussian random vectors with zero mean. We define a random vector $\boldsymbol{U}(t) = (U_1(t), \dots, U_n(t))$, $t \in [0,1]$, with

$$U_i(t) = \sqrt{t}X_i + \sqrt{1-t}Y_i,\ 1 \leq i \leq n.$$

Then the derivative of the function

$$f(t) = \mathbb{E}\left[F(\boldsymbol{U}(t))\right]$$

is

$$f'(t) = \sum_{i=1}^{n} \mathbb{E}\left[U_i'(t)\frac{\vartheta F}{\vartheta x_i}(\boldsymbol{U}(t))\right].$$

Furthermore, if F is two times differentiable, with all of its second order partial derivatives to be of moderate growth then

$$f'(t) = \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}\left(\mathbb{E}\left[X_iX_j\right] - \mathbb{E}\left[Y_iY_j\right]\right)\mathbb{E}\left[\frac{\vartheta^2 F}{\vartheta x_i \vartheta x_j}(\boldsymbol{U}(t))\right].$$

Proof. We want to use proposition 4.7.2 in order to find the derivative of f. The first thing we must do is establish condition (4.14) for F. So, let $I = [a,b] \subset [0,1]$ be an arbitrary compact subinterval. Then

$$\mathbb{E}\left[\sup_{t\in I}\left(F'(\boldsymbol{U}(t))\right)\right] = \mathbb{E}\left[\sup_{t\in I}\left(\frac{d}{dt}F(\boldsymbol{U}(t))\right)\right] = \mathbb{E}\left[\sup_{t\in I}\left(\sum_{i=1}^{n}\frac{\vartheta F(\boldsymbol{U}(t))}{\vartheta x_i}\frac{dx_i}{dt}\right)\right] \leq$$
$$\leq \sum_{i=1}^{n}\mathbb{E}\left[\sup_{t\in I}\left(U_i'(t)\frac{\vartheta F}{\vartheta x_i}(\boldsymbol{U}(t))\right)\right].$$

Also, it is straightforward that

$$U_i'(t) = \frac{dU_i}{dt}(t) = \frac{1}{2\sqrt{t}}X_i - \frac{1}{2\sqrt{1-t}}Y_i,\ 1 \leq i \leq n.$$

As a result, it suffices to bound the expectation of an arbitrary term of the above sum.

$$\mathbb{E}\left[\sup_{t\in I}\left|U_i'(t)\frac{\vartheta F}{\vartheta x_i}(\boldsymbol{U}(t))\right|\right] \leq \mathbb{E}\left[\sup_{t\in I}|U_i'(t)|\sup_{t\in I}\left|\frac{\vartheta F}{\vartheta x_i}(\boldsymbol{U}(t))\right|\right] \leq$$
$$\leq \sqrt{\mathbb{E}\left[\left(\sup_{t\in I}|U_i'(t)|\right)^2\right]}\cdot\sqrt{\mathbb{E}\left[\left(\sup_{t\in I}\left|\frac{\vartheta F}{\vartheta x_i}(\boldsymbol{U}(t))\right|\right)^2\right]} = \quad \text{(pro. (4.1.3))}$$
$$= \sqrt{\mathbb{E}\left[\sup_{t\in I}|U_i'(t)|^2\right]}\cdot\sqrt{\mathbb{E}\left[\sup_{t\in I}\left|\frac{\vartheta F}{\vartheta x_i}(\boldsymbol{U}(t))\right|^2\right]}.$$

We want to prove that the previous expression is bounded. We work with each term separately. For the first term we have that

$$\sqrt{\mathbb{E}\left[\sup_{t\in I}|U_i'(t)|^2\right]} = \sqrt{\mathbb{E}\left[\sup_{t\in I}\left(\frac{1}{4t}X_i^2 - \frac{1}{2\sqrt{t}\sqrt{1-t}}X_iY_i + \frac{1}{4(1-t)}Y_i^2\right)\right]} \leq$$

$$\leq \sqrt{\mathbb{E}\left[\sup_{t\in I}\left(\frac{1}{4t}X_i^2\right) + \sup_{t\in I}\left(\frac{1}{2\sqrt{t}\sqrt{1-t}}X_iY_i\right) + \sup_{t\in I}\left(\frac{1}{4(1-t)}Y_i^2\right)\right]} =$$

$$= \sqrt{\mathbb{E}\left[\frac{1}{4a}X_i^2\right] + \mathbb{E}\left[\frac{1}{4(1-b)}Y_i^2\right]} \leq \qquad (X_i, Y_i \text{ zero-mean, independent})$$

$$= \sqrt{\mathbb{E}\left[\left(\frac{1}{2\sqrt{a}}X_i\right)^2 + \left(\frac{1}{2\sqrt{1-b}}Y_i\right)^2\right]} \leq$$

$$\leq \sqrt{\mathbb{E}\left[\left(\frac{1}{2\sqrt{a}}X_i + \frac{1}{2\sqrt{1-b}}Y_i\right)^2\right]} \leq \qquad (\text{Triangle inequality, (4.1.1)}))$$

$$\leq \sqrt{\mathbb{E}\left[\frac{1}{4a}X_i^2\right]} + \sqrt{\mathbb{E}\left[\frac{1}{4(1-b)}Y_i^2\right]} \leq \infty$$

Now we move to the second term. We know that the first-order partial derivatives of F are of moderate growth. Thus, for every $c > 0$, there exists $A > 0$ such that

$$\left|\frac{\vartheta F}{\vartheta x_i}(\boldsymbol{x})\right| \leq Ae^{c\|\boldsymbol{x}\|_2^2}, \forall\, \boldsymbol{x} \in \mathbb{R}^n,\ 1 \leq i \leq n \tag{4.15}$$

Also,

$$\begin{aligned}\|\boldsymbol{U}(t)\|_2 &= \|\left(\sqrt{t}X_1 + \sqrt{1-t}Y_1, \ldots, \sqrt{t}X_n + \sqrt{1-t}Y_n\right)\|_2 = \\ &= \|\sqrt{t}\,(X_1, \ldots, X_n) + \sqrt{1-t}\,(Y_1, \ldots, Y_n)\|_2 \leq \\ &\leq \sqrt{t}\|\boldsymbol{X}\|_2 + \sqrt{1-t}\|\boldsymbol{Y}\|_2 \leq \\ &\leq 2max\left\{\|\boldsymbol{X}\|_2, \|\boldsymbol{Y}\|_2\right\}, \forall t \in I\end{aligned} \tag{4.16}$$

Combining (4.15), (4.16) yields

$$\sqrt{\mathbb{E}\left[\sup_{t\in I}\left|\frac{\vartheta F}{\vartheta x_i}(\boldsymbol{U}(t))\right|^2\right]} \leq \sqrt{\mathbb{E}\left[\sup_{t\in I}\left(Ae^{c\|\boldsymbol{U}(t)\|_2^2}\right)^2\right]} =$$

$$= \sqrt{\mathbb{E}\left[A^2\sup_{t\in I}\left(e^{2c\|\boldsymbol{U}(t)\|_2^2}\right)\right]} \leq$$

$$\leq A\sqrt{\mathbb{E}\left[e^{8c\,(max\{\|\boldsymbol{X}\|_2, \|\boldsymbol{Y}\|_2\})^2}\right]} = \qquad (eq.4.16)$$

$$= A\sqrt{\mathbb{E}\left[max\left\{e^{8c\|\boldsymbol{X}\|_2^2}, e^{8c\|\boldsymbol{Y}\|_2^2}\right\}\right]} \leq$$

$$\leq A\sqrt{\mathbb{E}\left[e^{8c\|\boldsymbol{X}\|_2^2} + e^{8c\|\boldsymbol{Y}\|_2^2}\right]}.$$

We know that $\boldsymbol{X}$ and $\boldsymbol{Y}$ are Gaussian vectors, with zero mean. Hence, we can write them as $\boldsymbol{X} = D\boldsymbol{g}_1$ and $\boldsymbol{Y} = D'\boldsymbol{g}_2$, where D, D' are $n \times n$ matrices and $\boldsymbol{g}_1, \boldsymbol{g}_2 \in \mathbb{R}^n$ are independent standard Gaussian random vectors. Then,

$$\begin{aligned}
\sqrt{\mathbb{E}\left[\sup_{t\in I}\left|\frac{\vartheta F}{\vartheta x_i}(\boldsymbol{U}(t))\right|^2\right]} &\leq A\sqrt{\mathbb{E}\left[e^{8c\|D\|_{2\to2}^2\|\boldsymbol{g}_1\|_2^2+8c\|D'\|_{2\to2}^2\|\boldsymbol{g}_2\|_2^2}\right]} = \quad (\|\boldsymbol{X}\|_2^2 \leq \|D\|_{2\to2}^2\|\boldsymbol{g}_1\|_2^2) \\
&= A\sqrt{\mathbb{E}\left[\prod_{i=1}^n e^{8c\|D\|_{2\to2}^2(g_1)_i^2}\cdot\prod_{i=1}^n e^{8c\|D'\|_{2\to2}^2(g_2)_i^2}\right]} = \\
&= A\sqrt{\mathbb{E}\left[\prod_{i=1}^n e^{8c\|D\|_{2\to2}^2(g_1)_i^2}\right]\cdot\mathbb{E}\left[\prod_{i=1}^n e^{8c\|D'\|_{2\to2}^2(g_2)_i^2}\right]} = \quad \text{(Independence)} \\
&= A\sqrt{\prod_{i=1}^n\mathbb{E}\left[e^{8c\|D\|_{2\to2}^2(g_1)_i^2}\right]\cdot\prod_{i=1}^n\mathbb{E}\left[e^{8c\|D'\|_{2\to2}^2(g_2)_i^2}\right]} = \text{(Lemma 4.2.1)} \\
&= A\sqrt{\prod_{i=1}^n\left(\frac{1}{\sqrt{1-16c\|D\|_{2\to2}^2}}e^0\right)\cdot\prod_{i=1}^n\left(\frac{1}{\sqrt{1-16c\|D'\|_{2\to2}^2}}e^0\right)} = \\
&= A\sqrt{\left(\frac{1}{\sqrt{1-16c\|D\|_{2\to2}^2}}\right)^n\cdot\left(\frac{1}{\sqrt{1-16c\|D'\|_{2\to2}^2}}\right)^n}.
\end{aligned}$$

The usage of lemma 4.2.1 is permitted, since we can choose c such that, $c = min\left\{\frac{1}{16\|D\|_{2\to2}^2}, \frac{1}{16\|D\|_{2\to2}^2}\right\}$.

As a result,

$$\sqrt{\mathbb{E}\left[\sup_{t\in I}\left|\frac{\vartheta F}{\vartheta x_i}(\boldsymbol{U}(t))\right|^2\right]} = A\sqrt{\left(\frac{1}{\sqrt{1-16c\|D\|_{2\to2}^2}}\right)^n\cdot\left(\frac{1}{\sqrt{1-16c\|D'\|_{2\to2}^2}}\right)^n} < \infty$$

and consequently

$$\mathbb{E}\left[\sup_{t\in I}\left|U_i'(t)\frac{\vartheta F}{\vartheta x_i}(\boldsymbol{U}(t))\right|\right] \leq \infty$$

Then from proposition (4.7.2) we have that

$$f'(t) = \mathbb{E}\left[\frac{d}{dt}F(\boldsymbol{U}(t))\right] = \sum_{i=1}^n\mathbb{E}\left[U_i'(t)\frac{\vartheta F}{\vartheta x_i}(\boldsymbol{U}(t))\right].$$

Furthermore, if F is twice differentiable with all of its second-order partial derivatives of moderate growth we can use proposition 4.7.1, i.e. Stein's lemma and obtain

$$f'(t) = \sum_{i=1}^n\mathbb{E}\left[U_i'(t)\frac{\vartheta F}{\vartheta x_i}(\boldsymbol{U}(t))\right] = \sum_{i=1}^n\sum_{j=1}^n\mathbb{E}\left[U_i'(t)U_j(t)\right]\mathbb{E}\left[\frac{\vartheta^2 F}{\vartheta x_i\vartheta x_j}(\boldsymbol{U}(t))\right].$$

Working the first expression inside the summation yields

$$\begin{aligned}\mathbb{E}\left[U_i'(t)U_j(t)\right] &= \mathbb{E}\left[\sqrt{t}\frac{1}{2\sqrt{t}}X_jX_i - \sqrt{t}\frac{1}{2\sqrt{1-t}}X_jY_i + \frac{1}{2\sqrt{t}}\sqrt{1-t}Y_jX_i - \frac{1}{2\sqrt{1-t}}\sqrt{1-t}\frac{1}{2\sqrt{t}}Y_jY_i\right] \\ &= \frac{1}{2}\left(\mathbb{E}\left[X_iX_j\right] - \mathbb{E}\left[Y_iY_j\right]\right).\end{aligned}$$

.

This leads to the final form of the derivative of f,

$$f'(t) = \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}\left(\mathbb{E}\left[X_iX_j\right] - \mathbb{E}\left[Y_iY_j\right]\right)\mathbb{E}\left[\frac{\vartheta^2 F}{\vartheta x_i \vartheta x_j}(\boldsymbol{U}(t))\right].$$

□

Lemma 4.7.1. *Let $F : \mathbb{R}^n \to \mathbb{R}$ be a Lipschitz function for some constant $L > 0$. Also, let $\boldsymbol{X} = (X_1, \ldots, X_n)$ and $\boldsymbol{Y} = (Y_1, \ldots, Y_n)$ be two Gaussian random vectors, with zero mean. Suppose that (in the distributional sense)*

$$\left(\mathbb{E}\left[|X_i - X_j|^2\right] - \mathbb{E}\left[|Y_i - Y_j|^2\right]\right)\frac{\vartheta^2 F}{\vartheta x_i \vartheta x_j} \geq 0,\ 1 \leq i, j \leq n \tag{4.17}$$

and

$$F(\boldsymbol{x} + t\boldsymbol{e}) = F(\boldsymbol{x}) + ct,\ \forall\, \boldsymbol{x} \in \mathbb{R}^n, \tag{4.18}$$

where c is some constant and $\boldsymbol{e} \in \mathbb{R}^n$, with $\boldsymbol{e} = (1, \ldots, 1)$. Then

$$\mathbb{E}\left[F(\boldsymbol{X})\right] \leq \mathbb{E}\left[F(\boldsymbol{Y})\right] \tag{4.19}$$

Proof. F is a Lipschitz function, so it holds that

$$\begin{aligned}|F(\boldsymbol{x}) - F(\boldsymbol{0})| \leq L\|\boldsymbol{x} - \boldsymbol{0}\|_2 \Rightarrow ||F(\boldsymbol{x})| - |F(\boldsymbol{0})|| \leq L\|\boldsymbol{x}\|_2 \Rightarrow \\ |F(\boldsymbol{x})| - |F(\boldsymbol{0})| \leq L\|\boldsymbol{x}\|_2 \Rightarrow \\ |F(\boldsymbol{x})| \leq L\|\boldsymbol{x}\|_2 + |F(\boldsymbol{0})|, \forall\, \boldsymbol{x} \in \mathbb{R}^n.\end{aligned}$$

As a result, we can establish that F is of moderate growth. Suppose $b > 0$, then

$$\lim_{\|\boldsymbol{x}\|_2 \to +\infty} F(\boldsymbol{x})e^{-b\|\boldsymbol{x}\|_2^2} \leq \lim_{\|\boldsymbol{x}\|_2 \to +\infty} \left(L\|\boldsymbol{x}\|_2 + |F(\boldsymbol{0})|\right)e^{-b\|\boldsymbol{x}\|_2^2} = 0.$$

First, we are going to prove the theorem for the case where F is a two times continuously differentiable function, whose first and second-order partial derivatives are of moderate growth.

Also, note that condition (4.18) is equivalent to the following expression

$$\sum_{j=1}^{n}\frac{\vartheta F}{\vartheta x_i \vartheta x_j}(\boldsymbol{x}) = 0, \forall\, \boldsymbol{x} \in \mathbb{R}^n,\ 1 \leq i \leq n.$$

Then,

$$\begin{aligned}\sum_{j=1}^{n}\frac{\vartheta F}{\vartheta x_i \vartheta x_j}(\boldsymbol{x}) = 0 \Rightarrow \sum_{j=1, j\neq i}^{n}\frac{\vartheta^2 F}{\vartheta x_i \vartheta x_j}(\boldsymbol{x}) + \frac{\vartheta^2 F}{\vartheta x_i^2}(\boldsymbol{x}) = 0 \Rightarrow \\ \frac{\vartheta^2 F}{\vartheta x_i^2}(\boldsymbol{x}) = -\sum_{j=1, j\neq i}^{n}\frac{\vartheta^2 F}{\vartheta x_i \vartheta x_j}(\boldsymbol{x}),\ \boldsymbol{x} \in \mathbb{R}^n,\ 1 \leq i \leq n.\end{aligned} \tag{4.20}$$

Consider the function

$$f(t) = \mathbb{E}\left[F(\boldsymbol{U}(t))\right], \, t \in [0,1].$$

which is defined as the corresponding function in proposition (4.7.3). As we said before, we deal with the case where F is a two times continuously differentiable function , whose partial derivatives up to second-order are of moderate growth. Hence, from proposition (4.7.3) the derivative of f is

$$f'(t) = \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}\left(\mathbb{E}\left[X_iX_j\right] - \mathbb{E}\left[Y_iY_j\right]\right)\mathbb{E}\left[\frac{\vartheta^2 F}{\vartheta x_i \vartheta x_j}(\boldsymbol{U}(t))\right], \, t \in [0,1].$$

Working the right-hand side of the previous expression for an arbitrary $\boldsymbol{x} \in \mathbb{R}^n$ yields

$$\begin{aligned}
&\sum_{i=1}^{n}\sum_{j=1}^{n}\left(\mathbb{E}\left[X_iX_j\right] - \mathbb{E}\left[Y_iY_j\right]\right)\frac{\vartheta^2 F}{\vartheta x_i \vartheta x_j}(\boldsymbol{x}) = \\
&= \sum_{i=1}^{n}\left[\left(\mathbb{E}\left[X_i^2\right] - \mathbb{E}\left[Y_i^2\right]\right)\frac{\vartheta^2 F}{\vartheta x_i^2}(\boldsymbol{x}) + \sum_{j=1, i\neq j}^{n}\left(\mathbb{E}\left[X_iX_j\right] - \mathbb{E}\left[Y_iY_j\right]\right)\frac{\vartheta^2 F}{\vartheta x_i x_j}(\boldsymbol{x})\right] = \, (eq.\,(4.20)) \\
&= -\sum_{i=1}^{n}\left[\left(\mathbb{E}\left[X_i^2\right] - \mathbb{E}\left[Y_i^2\right]\right)\sum_{j=1, i\neq j}^{n}\frac{\vartheta^2 F}{\vartheta x_i x_j}(\boldsymbol{x}) + \sum_{j=1, i\neq j}^{n}\left(\mathbb{E}\left[X_iX_j\right] - \mathbb{E}\left[Y_iY_j\right]\right)\frac{\vartheta^2 F}{\vartheta x_i x_j}(\boldsymbol{x})\right] = \\
&= -\frac{1}{2}\sum_{i=1}^{n}\sum_{j=1, i\neq j}^{n}\left[\left(\mathbb{E}\left[X_i^2\right] - \mathbb{E}\left[Y_i^2\right] + \mathbb{E}\left[X_j^2\right] - \mathbb{E}\left[Y_j^2\right] - 2\left(\mathbb{E}\left[X_iX_j\right] - \mathbb{E}\left[Y_iY_j\right]\right)\right)\frac{\vartheta^2 F}{\vartheta x_i x_j}(\boldsymbol{x})\right] \\
&= -\frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}\left[\left(\mathbb{E}\left[|X_i - X_j|^2\right] - \mathbb{E}\left[|Y_i - Y_j|^2\right]\right)\frac{\vartheta^2 F}{\vartheta x_i \vartheta x_j}(\boldsymbol{x})\right].
\end{aligned}$$

Condition (4.17) establishes that

$$-\frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}\left[\left(\mathbb{E}\left[|X_i - X_j|^2\right] - \mathbb{E}\left[|Y_i - Y_j|^2\right]\right)\frac{\vartheta^2 F}{\vartheta x_i \vartheta x_j}(\boldsymbol{x})\right] \leq 0$$

and consequently

$$\sum_{i=1}^{n}\sum_{j=1}^{n}\left(\mathbb{E}\left[X_iX_j\right] - \mathbb{E}\left[Y_iY_j\right]\right)\frac{\vartheta^2 F}{\vartheta x_i \vartheta x_j} \leq 0, \, \forall \, \boldsymbol{x} \in \mathbb{R}^n.$$

It is not difficult to see that

$$f'(t) \leq 0, \, t \in [0,1].$$

That means that f is a decreasing function in $[0,1]$ and thus,

$$f(1) \leq f(0) \Rightarrow \mathbb{E}\left[f(\boldsymbol{X})\right] \leq \mathbb{E}\left[f(\boldsymbol{Y})\right].$$

Turning now to the general case, where there is no guarantee that F is two times continuously differentiable, we seek to find a two times continuously differentiable approximation of F.

Let $\psi(\boldsymbol{x})$ be a two times continuously differentiable, nonnegative function, with support in $\mathcal{B}(\mathbf{0}, 1)$, such that

$$\int_{\mathbb{R}^n} \psi(\boldsymbol{x}) d\boldsymbol{x} = 1.$$

Also, we define the following function

$$\psi_h = h^{-n}\psi(\frac{\boldsymbol{x}}{h}),\ h > 0,$$

for which it holds that

$$\int_{\mathbb{R}^n} \psi_h(\boldsymbol{x}) d\boldsymbol{x} = \int_{\mathbb{R}^n} h^{-n}\psi(\frac{\boldsymbol{x}}{h}) d\boldsymbol{x} = h^{-n} \int_{\mathbb{R}^n} \psi(\boldsymbol{y}) h^n d\boldsymbol{y} = 1.$$

It is easy to see that

$$supp(\psi) \subseteq \mathcal{B}(0,1) \Rightarrow supp(\psi_h) \subseteq \mathcal{B}(0,h).$$

We define the following sequence of functions

$$F_h(\boldsymbol{x}) = (F * \psi_h)(\boldsymbol{x}) = \int_{\mathbb{R}^n} F(\boldsymbol{y})\psi_h(\boldsymbol{x}-\boldsymbol{y}) d\boldsymbol{y} = \int_{\mathbb{R}^n} F(\boldsymbol{x}-\boldsymbol{y})\psi_h(\boldsymbol{y}) d\boldsymbol{y}. \quad (4.21)$$

Furthermore, we can see that Lebesgue's dominated convergence theorem (theorem A.3.1) allows us to interchange the derivative with the integral. Therefore, we can easily see that F_h is two times continuously differentiable. Indeed,

$$\frac{\vartheta F_h}{\vartheta x_i}(\boldsymbol{x}) = \frac{\vartheta}{\vartheta x_i} \int_{\mathbb{R}^n} F(\boldsymbol{y})\psi_h(\boldsymbol{x}-\boldsymbol{y}) d\boldsymbol{y} = \int_{\mathbb{R}^n} F(\boldsymbol{y}) \frac{\vartheta \psi_h}{\vartheta x_i}(\boldsymbol{x}-\boldsymbol{y}) d\boldsymbol{y} = F * \frac{\vartheta \psi_h}{\vartheta x_i}.$$

In the same way we can also attain

$$\frac{\vartheta^2 F_h}{\vartheta x_i \vartheta x_j}(\boldsymbol{x}) = F * \frac{\vartheta^2 \psi_h}{\vartheta x_i \vartheta x_j}.$$

Also, it is easy to verify that the first and second order partial derivatives of F_h are of moderate growth, using the definition of convolution and the fact that ψ_h has compact support and is two times continuously differentiable.

We know that, in the distributional sense, it holds that $\dfrac{\vartheta^2 F}{\vartheta x_i \vartheta x_j} \geq 0$, if for all nonnegative functions g, two times differentiable, with compact support we have that

$$\int_{\mathbb{R}^n} F(\boldsymbol{x}) \frac{\vartheta^2 g}{\vartheta x_i \vartheta x_j}(\boldsymbol{x}) d\boldsymbol{x} \geq 0.$$

So, let g be a nonnegative, two times continuously differentiable function on $\mathbb{R}^n$, with compact support. Then,

$$\begin{aligned}
\int_{\mathbb{R}^n} F_h(\boldsymbol{x})\frac{\vartheta^2 g}{\vartheta x_i \vartheta x_j}(\boldsymbol{x})d\boldsymbol{x} &= \int_{\mathbb{R}^n}\left(\int_{\mathbb{R}^n} F(\boldsymbol{y})\psi_h(\boldsymbol{x}-\boldsymbol{y})d\boldsymbol{y}\right)\frac{\vartheta^2 g}{\vartheta x_i \vartheta x_j}(\boldsymbol{x})d\boldsymbol{x} = && \text{(Fubini's theorem)}\\
&= \int_{\mathbb{R}^n} F(\boldsymbol{y})\left(\int_{\mathbb{R}^n}\psi_h(\boldsymbol{y}-\boldsymbol{x})\frac{\vartheta^2 g}{\vartheta x_i \vartheta x_j}(\boldsymbol{x})d\boldsymbol{x}\right)d\boldsymbol{y} = && \text{(Lebesgue's th. (A.3.1))}\\
&= \int_{\mathbb{R}^n} F(\boldsymbol{y})\frac{\vartheta^2}{\vartheta x_i \vartheta x_j}\left(\int_{\mathbb{R}^n}\psi_h(\boldsymbol{y}-\boldsymbol{x})g(\boldsymbol{x})d\boldsymbol{x}\right)d\boldsymbol{y} = \\
&= \int_{\mathbb{R}^n} F(\boldsymbol{y})\frac{\vartheta^2}{\vartheta x_i \vartheta x_j}(\psi_h * g)(\boldsymbol{y})d\boldsymbol{y}.
\end{aligned}$$

As a result, condition (4.17) holds for F_h.

Moving to establishing the condition (4.18) for F_h, we can write that

$$\begin{aligned}
F_h(\boldsymbol{x}+t\boldsymbol{e}) &= (F*\psi_h)(\boldsymbol{x}+t\boldsymbol{e}) = \psi_h(\boldsymbol{x}+t\boldsymbol{e}) * F = \\
&= \int_{\mathbb{R}^n} F(\boldsymbol{x}+t\boldsymbol{e}-\boldsymbol{y})\psi_h(\boldsymbol{y})d\boldsymbol{y} = \\
&= \int_{\mathbb{R}^n} (F(\boldsymbol{x}-\boldsymbol{y})+ct)\psi_h(\boldsymbol{y})d\boldsymbol{y} = \\
&= \int_{\mathbb{R}^n} F(\boldsymbol{x}-\boldsymbol{y})\psi_h(\boldsymbol{y})d\boldsymbol{y} + \int_{\mathbb{R}^n} ct\psi_h(\boldsymbol{y})d\boldsymbol{y} = \\
&= F_h(\boldsymbol{x}) + ct\int_{\mathbb{R}^n}\psi_h(\boldsymbol{x})d\boldsymbol{x} = \\
&= F_h(\boldsymbol{x}) + ct.
\end{aligned}$$

Consequently, the function F_h satisfies all the conditions of this theorem and in addition the conditions imposed by us in the first half of this proof. Using the results obtained on the first half of this proof we can write that

$$\mathbb{E}\left[F_h(\boldsymbol{X})\right] \leq \mathbb{E}\left[F_h(\boldsymbol{Y})\right], \forall h > 0.$$

Moreover, we want to show that F_h converges uniformly to F. Before we do that notice that

$$\boldsymbol{y} \in \mathcal{B}(\boldsymbol{x}, h) \Rightarrow \|\boldsymbol{x}-\boldsymbol{y}\|_2 \leq h \Rightarrow \boldsymbol{x}-\boldsymbol{y} \in \mathcal{B}(\boldsymbol{0}, h) \supseteq supp(\psi_h).$$

In order to prove the uniform convergence of F_h to F we consider the following sum.

$$
\begin{aligned}
|F_h(\boldsymbol{x}) - F(\boldsymbol{x})| &= \left| \int_{\mathbb{R}^n} (F(\boldsymbol{y}) - F(\boldsymbol{x}))\psi_h(\boldsymbol{x}-\boldsymbol{y})d\boldsymbol{y} \right| \leq \\
&\leq \int_{\mathcal{B}(\boldsymbol{x},h)} |F(\boldsymbol{y}) - F(\boldsymbol{x})|\, \psi_h(\boldsymbol{x}-\boldsymbol{y})d\boldsymbol{y} \leq \quad (F \text{ Lipschitz}), \quad (\psi_h \text{ nonnegative}) \\
&\leq \int_{\mathcal{B}(\boldsymbol{x},h)} L\|\boldsymbol{y}-\boldsymbol{x}\|_2 \psi_h(\boldsymbol{x}-\boldsymbol{y})d\boldsymbol{y} \leq \\
&\leq \int_{\mathcal{B}(\boldsymbol{x},h)} Lh\psi_h(\boldsymbol{x}-\boldsymbol{y})d\boldsymbol{y} \leq \\
&\leq Lh, \forall \boldsymbol{x} \in \mathbb{R}^n.
\end{aligned}
$$

Therefore, we have established that for $h \to 0$, F_h converges uniformly to F. Thus, the uniform convergence of F_h can be exploited to show that

$$
\begin{aligned}
\mathbb{E}\left[F(\boldsymbol{X})\right] = \mathbb{E}\left[\lim_{h\to 0} F_h(\boldsymbol{X})\right] &= \lim_{h\to 0} \mathbb{E}\left[F_h(\boldsymbol{X})\right] \leq \\
&\leq \lim_{h\to 0} \mathbb{E}\left[F_h(\boldsymbol{Y})\right] = \mathbb{E}\left[\lim_{h\to 0} F_h(\boldsymbol{Y})\right] = \mathbb{E}\left[F(\boldsymbol{Y})\right].
\end{aligned}
$$

□

At this point we have all the tools at our disposal that are necessary in order to prove Gordon's Lemma .

Theorem 4.7.1 (Gordon's Lemma)**.** *Let $X_{i,j}$, $Y_{i,j}$, $1 \leq i \leq n$, $1 \leq j \leq m$ be two finite families of Gaussian random variables with zero mean. If*

$$
\mathbb{E}\left[|X_{i,j} - X_{k,l}|^2\right] \leq \mathbb{E}\left[|Y_{i,j} - Y_{k,l}|^2\right], \forall i \neq k, j, l
$$

$$
\mathbb{E}\left[|X_{i,j} - X_{i,l}|^2\right] \geq \mathbb{E}\left[|Y_{i,j} - Y_{i,l}|^2\right], \forall i, j, l,
$$

then it holds that

$$
\mathbb{E}\left[\min_{1\leq i\leq n} \max_{1\leq j\leq m} X_{i,j}\right] \geq \mathbb{E}\left[\min_{1\leq i\leq n} \max_{1\leq j\leq m} Y_{i,j}\right].
$$

Proof. We define the following function

$$
F(\boldsymbol{x}) = \min_{1\leq i\leq n} \max_{1\leq j\leq m} x_{ij} \tag{4.22}
$$

where $\boldsymbol{x} \in \mathbb{R}^{nm}$ is a doubly-indexed vector $\boldsymbol{x} = (x_{ij})_{1\leq i\leq n, 1\leq j\leq m}$.

We are going to use lemma 4.7.1. We can easily see that F is Lipschitz, with constant 1. Indeed,

$$
\begin{aligned}
|F(\boldsymbol{x}) - F(\boldsymbol{y})| &= \left| \min_{1\leq i\leq n} \max_{1\leq j\leq m} x_{ij} - \min_{1\leq i\leq n} \max_{1\leq j\leq m} y_{ij} \right| = |x_{i_1 j_1} - y_{i_2 j_2}| \leq \\
&\leq \sqrt{\sum_{i=1}^{n}\sum_{j=1}^{m} |x_{ij} - y_{ij}|^2} = \|\boldsymbol{x} - \boldsymbol{y}\|_2
\end{aligned}
$$

Next we want to show that (4.17) holds. We notice that only two variables are involved each time, so we can choose two of the variables and fix the others. Let $t = x_{ij}$ and $s = x_{kl}$. Then, we can express F as

$$F(\boldsymbol{x}) = \begin{cases} A(t,s) & ,i = k \\ B(t,s) & ,i \neq k \end{cases} = \begin{cases} max\left\{f(t), g(s)\right\} & ,i = k \\ min\left\{f(t), g(s)\right\} & ,i \neq k \end{cases}.$$

The functions f, g are of the form

$$l(t) = \begin{cases} a & ,t < a \\ t & ,a \leq t \leq b \\ b & ,t > b \end{cases}, \tag{4.23}$$

where $a \leq b$. Note that it is possible that $a = -\infty, b = +\infty$.
The weak derivative of l is

$$l'(t) = \begin{cases} 0 & ,t < a \\ 1 & ,a \leq t \leq b \\ 0 & ,t > b \end{cases} \tag{4.24}$$

The function $A(t,s)$ can be equivalently written as

$$A(t,s) = max\left\{f(t), g(s)\right\} = \frac{1}{2}\left(f(t) + g(s) + |f(t) - g(s)|\right).$$

We calculate the partial weak derivative of A with respect to t.

$$h(t,s) = \frac{\vartheta}{\vartheta t}A(t,s) = \frac{1}{2}\left(f'(t) + f'(t)sgn(f(t) - g(s))\right) = \\ = \begin{cases} \frac{1}{2}\left(1 + sgn(t - g(s))\right) & ,a \leq t \leq b \\ 0 & ,a < t, t > b \end{cases}.$$

The function $h(t,s)$, with fixed t, is nonincreasing, thus $\frac{\vartheta^2}{\vartheta s \vartheta t}A(t,s) \leq 0$, in the distributional sense.
Moving to the other case we have

$$B(t,s) = min\left\{f(t), g(s)\right\} = \frac{1}{2}\left(f(t) + g(s) - |f(t) - g(s)|\right).$$

Using the same reasoning with the previous case we can conclude that $\frac{\vartheta^2}{\vartheta s \vartheta t}B(t,s) \geq 0$, in the distributional sense.
Consequently, in the sense of distributional derivatives it holds that

$$\begin{cases} \frac{\vartheta^2 F}{\vartheta x_{ij}\vartheta x_{kl}} \leq 0 & ,i = k \\ \frac{\vartheta^2 F}{\vartheta x_{ij}\vartheta x_{kl}} \geq 0 & ,i \neq k \end{cases}.$$

Furthermore, the following condition is provided in the definition of Gordon's lemma,

$$\begin{cases} \mathbb{E}\left[|X_{i,j}-X_{k,l}|^2\right]-\mathbb{E}\left[|Y_{i,j}-Y_{k,l}|^2\right]\geq 0 & ,i=k \\ \mathbb{E}\left[|X_{i,j}-X_{k,l}|^2\right]-\mathbb{E}\left[|Y_{i,j}-Y_{k,l}|^2\right]\leq 0 & ,i\neq k \end{cases}. \tag{4.25}$$

Thus,

$$\begin{cases} \dfrac{\vartheta^2 F}{\vartheta x_{ij}\vartheta x_{kl}}\leq 0 & ,i=k \\ \dfrac{\vartheta^2 F}{\vartheta x_{ij}\vartheta x_{kl}}\geq 0 & ,i\neq k \end{cases}\cdot\begin{cases} \mathbb{E}\left[|X_{i,j}-X_{k,l}|^2\right]-\mathbb{E}\left[|Y_{i,j}-Y_{k,l}|^2\right]\geq 0 & ,i=k \\ \mathbb{E}\left[|X_{i,j}-X_{k,l}|^2\right]-\mathbb{E}\left[|Y_{i,j}-Y_{k,l}|^2\right]\leq 0 & ,i\neq k \end{cases}=$$

$$=\left(\mathbb{E}\left[|X_{i,j}-X_{k,l}|^2\right]-\mathbb{E}\left[|Y_{i,j}-Y_{k,l}|^2\right]\right)\cdot\frac{\vartheta^2 F}{\vartheta x_{ij}\vartheta x_{kl}}\leq 0,\forall\, i,j,k,l.$$

In addition, for $r\in\mathbb{R}$

$$F(\boldsymbol{x}+r\boldsymbol{e})=\min_{1\leq i\leq n}\max_{1\leq j\leq m}(x_{ij}+r)=\min_{1\leq i\leq n}\max_{1\leq j\leq m}(x_{ij})+r=F(\boldsymbol{x})+r.$$

Then, the conditions of lemma (4.7.1) are satisfied for $-F$ and as a result it holds that

$$\mathbb{E}\left[-F(\boldsymbol{X})\right]\leq\mathbb{E}\left[-F(\boldsymbol{Y})\right]\Rightarrow\mathbb{E}\left[F(\boldsymbol{X})\right]\geq\mathbb{E}\left[F(\boldsymbol{Y})\right] \tag{4.26}$$

□

An important observation is the fact that Gordon's lemma also holds for Gaussian processes. This result is a combination of two things. First, we know that any finite subset of a Gaussian process is a Gaussian random vector, i.e. if $(X_i)_{i\in T}$ is a Gaussian stochastic process, then the vector $(X_i)_{i\in S}$, $S\subseteq T$ with $card(S)<\infty$, is a Gaussian random vector. Secondly, the suprema we consider here are lattice suprema, which roughly means that the supremum is computed on finite subsets of the stochastic process.

An immediate consequence of Gordon's Lemma is *Slepian's Lemma.*

Corollary 4.7.1 (Slepian's Lemma)**.** *Let $\boldsymbol{X},\boldsymbol{Y}\in\mathbb{R}^n$ be two Gaussian random vectors with zero mean. If*

$$\mathbb{E}\left[|X_i-X_j|^2\right]\leq\mathbb{E}\left[|Y_i-Y_j|^2\right],\ 1\leq i,j\leq n, \tag{4.27}$$

then it holds that

$$\mathbb{E}\left[\max_{1\leq i\leq n}X_i\right]\leq\mathbb{E}\left[\max_{1\leq i\leq n}Y_i\right]. \tag{4.28}$$

4.8 Concentration of measure

In this section we aim at proving a concentration of measure argument. We are going to follow the entropy method. As with every complex argument the proof requires several intermediate propositions, definitions and lemmas.

First, we need to define the *entropy* of a random variable.

Definition 4.8.1 (Entropy)**.** *Let X be a nonnegative random variable on a probability space $(\Omega, \Sigma, \mathbb{P})$. Also, let define the convex function $\phi(x) = xln(x)$, $x > 0$, which we continuously extend to $x = 0$ by setting $\phi(x) = 0$. The entropy of X is defined as*

$$\mathcal{E}(X) = \mathbb{E}[\phi(X)] - \phi(\mathbb{E}[X]) = \mathbb{E}[Xln(X)] - \mathbb{E}[X]\, ln(\mathbb{E}[X]).$$

If $\mathbb{E}[Xln(X)] = \infty$, we set $\mathcal{E}(X) = \infty$.

We can easily verify that the entropy is a nonnegative quantity, i.e $\mathcal{E}(X) \geq 0$ and it is homogeneous, i.e $\mathcal{E}(tX) = t\mathcal{E}(X)$, $\forall\, t > 0$.

We are going to use the entropy in order to develop a concentration inequality for a random variable X. This approach is called the *entropy method*. Specifically, we want to obtain a bound of the form

$$\mathcal{E}\left(e^{tX}\right) \leq g(t)\mathbb{E}\left[e^{tX}\right],$$

for $t > 0$ and some suitable function $g(t)$. Then we can use the following procedure in order to obtain a concentration inequality for X.

First of all, note that $M_X'(t) = \mathbb{E}\left[Xe^{tX}\right]$. As a result, we can rewrite the entropy of e^{tX} as

$$\begin{aligned}\mathcal{E}(e^{tX}) &= tM_X'(t) - M_X(t)ln\left(M_X(t)\right) \leq g(t)M_X(t)\\ &\Rightarrow \frac{M_X'(t)}{tM_X(t)} - \frac{ln(M_X(t))}{t^2} \leq \frac{g(t)}{t^2}\end{aligned}$$

Then by setting $F(t) = \dfrac{ln(M_X(t))}{t}$ we can deduce that

$$F'(t) \leq \frac{g(t)}{t^2} \Rightarrow F(t) - F(0) \leq \int_0^t \frac{g(z)}{z^2}dz \Rightarrow \qquad (F(0) = \lim_{t\to 0}\frac{ln(M_X(t))}{t} = \mathbb{E}[X])$$

$$F(t) - \mathbb{E}[X] \leq \int_0^t \frac{g(z)}{z^2}dz \Rightarrow \frac{ln(M_X(t))}{t} - \mathbb{E}[X] \leq \int_0^t \frac{g(z)}{z^2}dz \Rightarrow$$

$$e^{ln(M_X(t)) - t\mathbb{E}[X]} \leq exp\left(t\int_0^t \frac{g(z)}{z^2}dz\right) \Rightarrow$$

$$\mathbb{E}\left[e^{t(X-\mathbb{E}[X])}\right] \leq exp\left(t\int_0^t \frac{g(z)}{z^2}dz\right).$$

Combining Markov's theorem with the previous result yields the following tail bound

$$\mathbb{P}[X - \mathbb{E}[X] \geq s] \leq \frac{exp\left(t\int_0^t \frac{g(z)}{z^2}dz\right)}{e^{ts}},\ s > 0$$

This approach is also called the *Herbst argument*. The next lemma contains some basic results about entropy.

Lemma 4.8.1 (Results about entropy)**.** *Let X be a nonnegative random variable such that $\mathbb{E}[X] < \infty$. Then, it holds that*

(a)

$$\mathcal{E}(X) = sup\left\{\mathbb{E}[XY] : \mathbb{E}\left[e^Y\right] \leq 1\right\}, \tag{4.29}$$

(b)

$$\mathcal{E}(X) = \sup_{Z>0}\left\{\mathbb{E}[Xln(Z)] - \mathbb{E}[X]\, ln(\mathbb{E}[Z])\right\}.$$

(c) The entropy is a subadditive function, i.e

$$\mathcal{E}(X+Z) \leq \mathcal{E}(X) + \mathcal{E}(Z),$$

where Z is also a nonnegative random variable.

Proof. (a) First of all, assume that X is a strictly positive random variable. Notice that the homogeneity of entropy allows us to assume that $\mathbb{E}[X] = 1$. Also, let Y be a random variable such that $\mathbb{E}\left[e^Y\right] \leq 1$. Using Fenchel's inequality (A.5.3) and the fact that $\mathbb{E}\left[e^Y\right] \leq 1$ we can obtain the following expression.

$$\mathbb{E}[XY] \leq \mathbb{E}\left[e^Y\right] + \mathbb{E}[Xln(X)] - exX \leq \mathbb{E}[Xln(X)] = \mathcal{E}(X).$$

Hence,

$$sup\left\{\mathbb{E}[XY] : \mathbb{E}\left[e^Y\right] \leq 1\right\} \leq \mathcal{E}(X).$$

In order to show the opposite inequality, we pick $Y = ln(X) - ln(\mathbb{E}[X])$. For that Y we can easily verify that

$$\mathbb{E}\left[e^Y\right] = \mathbb{E}\left[e^{ln(X)-ln(\mathbb{E}[X])}\right] = \mathbb{E}\left[X\frac{1}{\mathbb{E}[X]}\right] = 1.$$

Then,

$$\mathbb{E}[XY] = \mathbb{E}[Xln(X) - Xln(\mathbb{E}[X])] = \mathcal{E}(X).$$

Thus,

$$\mathcal{E}(X) \leq sup\left\{\mathbb{E}[XY] : \mathbb{E}\left[e^Y\right] \leq 1\right\},$$

and the proof for a strictly positive random variable is complete. It is now easy to show that the main result holds for a nonnegative random variable using the continuity of ϕ(the function introduced in the definition of entropy) at 0 and an elementary approximation argument.

(b) This result is obtained by setting in (4.29) $Y = ln\left(\frac{Z}{\mathbb{E}[Z]}\right)$, where Z is a positive random variable. Notice that $\mathbb{E}\left[e^Y\right] \leq 1$ is satisfied for every $Z > 0$.

(c) It is straightforward that

$$\begin{aligned}\mathcal{E}(X+Z) &= sup\left\{\mathbb{E}\left[(X+Z)Y\right] : \mathbb{E}\left[e^Y\right] \leq 1\right\} \leq \\ &\leq sup\left\{\mathbb{E}\left[XY\right] : \mathbb{E}\left[e^Y\right] \leq 1\right\} + sup\left\{\mathbb{E}\left[ZY\right] : \mathbb{E}\left[e^Y\right] \leq 1\right\} = \\ &= \mathcal{E}(X) + \mathcal{E}(Z).\end{aligned}$$

□

Before we proceed to the next proposition we need to give three more definitions. Let $\boldsymbol{X} = (X_1, X_2, \ldots, X_n)$ be a random vector and $f : \mathbb{R}^n \to \mathbb{R}$ be a function. We denote with $\boldsymbol{X}^{(i)}$ the following vector

$$\boldsymbol{X}^{(i)} = (X_1, \ldots, X_{i-1}, X_i, \ldots, X_n).$$

Now we can define the following quantities.

- Conditional expectation

$$\mathbb{E}_{X_i}\left[f(\boldsymbol{X})\right] = \mathbb{E}_{X_i}\left[f(X_1, \ldots, X_i, \ldots, X_n)\right] = \mathbb{E}\left[f(\boldsymbol{X})|\boldsymbol{X}^{(i)}\right].$$

  The conditional expectation is a function of $X_1, \ldots, X_{i-1}, X_{i+1}, \ldots, X_n$ and is constant with respect to X_i.

- Conditional entropy

$$\mathcal{E}_{X_i}\left[f(\boldsymbol{X})\right] = \mathcal{E}\left(f(\boldsymbol{X})|\boldsymbol{X}^{(i)}\right) = \mathbb{E}_{X_i}\left[f(\boldsymbol{X})ln(f(\boldsymbol{X}))\right] - \mathbb{E}_{X_i}\left[f(\boldsymbol{X})\right] ln\left(\mathbb{E}_{X_i}\left[f(\boldsymbol{X})\right]\right)$$

  The conditional entropy is a random variable that depends on $\boldsymbol{X}^{(i)}$.

- Conditional expectation operator

$$\mathbb{E}^i\left[f(\boldsymbol{X})\right] = \mathbb{E}_{X_1,\ldots,X_{i-1}}\left[f(\boldsymbol{X})\right] = \mathbb{E}\left[f(\boldsymbol{X})|X_i, \ldots, X_n\right]$$

  Notice that $\mathbb{E}^1\left[f(\boldsymbol{X})\right] = f(\boldsymbol{X})$ and $\mathbb{E}^{n+1}\left[f(\boldsymbol{X})\right] = \mathbb{E}\left[f(\boldsymbol{X})\right]$

We are now ready to prove the *tensorization inequality* for entropy.

Proposition 4.8.1 (Tensorization inequality)**.** *Let $\boldsymbol{X} = (X_1, \ldots, X_n)$ be a random vector, where X_i, $1 \leq i \leq n$, are independent random variables and f be a nonnegative function for which it holds that $\mathbb{E}\left[f(\boldsymbol{X})\right] < \infty$. Then*

$$\mathcal{E}(f(\boldsymbol{X})) \leq \mathbb{E}\left[\sum_{i=1}^{n} \mathcal{E}_{X_i}\left[f(\boldsymbol{X})\right]\right].$$

Proof. First, assume that f is strictly positive. The use of lemma 4.8.1(b) with $Z = \mathbb{E}^i\left[f(\boldsymbol{X})\right] > 0$ leads to

$$\begin{aligned}&\mathbb{E}_{X_i}\left[f(\boldsymbol{X})ln\left(\mathbb{E}^i\left[f(\boldsymbol{X})\right]\right)\right] - \mathbb{E}_{X_i}\left[f(\boldsymbol{X})\right] ln\left(\mathbb{E}_{X_i}\left[\mathbb{E}^i\left[f(\boldsymbol{X})\right]\right]\right) \leq \mathcal{E}_{X_i}\left[f(\boldsymbol{X})\right] \\ &\Rightarrow \mathbb{E}_{X_i}\left[f(\boldsymbol{X})ln\left(\mathbb{E}^i\left[f(\boldsymbol{X})\right]\right) - f(\boldsymbol{X})ln\left(\mathbb{E}_{X_i}\left[\mathbb{E}^i\left[f(\boldsymbol{X})\right]\right]\right)\right] \leq \mathcal{E}_{X_i}\left[f(\boldsymbol{X})\right].\end{aligned} \tag{4.30}$$

Using the fact that X_i, $1 \leq i \leq n$ are independent random variables and Fubini's theorem we can obtain the following result

$$\mathbb{E}_{X_i}\left[\mathbb{E}^i\left[f(\boldsymbol{X})\right]\right] = \mathbb{E}_{X_i}\left[\mathbb{E}_{X_1,\ldots,X_{i-1}}\left[f(\boldsymbol{X})\right]\right] = \mathbb{E}^{i+1}\left[f(\boldsymbol{X})\right]. \tag{4.31}$$

Also, we create the following telescopic sum

$$ln\left(f(\boldsymbol{X})\right) - ln\left(\mathbb{E}\left[f(\boldsymbol{X})\right]\right) = \sum_{i=1}^{n}\left(ln\left(\mathbb{E}^i\left[f(\boldsymbol{X})\right]\right) - ln\left(\mathbb{E}^{i+1}\left[f(\boldsymbol{X})\right]\right)\right),$$

using the conditional expectation operator. Then, using the above expression we can obtain the following

$$\begin{aligned}
\mathcal{E}(f(\boldsymbol{X})) &= \mathbb{E}\left[f(\boldsymbol{X})ln(f(\boldsymbol{X})) - f(\boldsymbol{X})ln\left(\mathbb{E}\left[f(\boldsymbol{X})\right]\right)\right] = \\
&= \mathbb{E}\left[f(\boldsymbol{X})\sum_{i=1}^{n}\left(ln\left(\mathbb{E}^i\left[f(\boldsymbol{X})\right]\right) - ln\left(\mathbb{E}^{i+1}\left[f(\boldsymbol{X})\right]\right)\right)\right] = \\
&= \sum_{i=1}^{n}\mathbb{E}\left[f(\boldsymbol{X})ln\left(\mathbb{E}^i\left[f(\boldsymbol{X})\right]\right) - f(\boldsymbol{X})ln\left(\mathbb{E}_{X_i}\left[\mathbb{E}^i\left[f(\boldsymbol{X})\right]\right]\right)\right] = \qquad (eq.(4.31)) \\
&= \sum_{i=1}^{n}\mathbb{E}\left[\mathbb{E}_{X_i}\left[f(\boldsymbol{X})ln\left(\mathbb{E}^i\left[f(\boldsymbol{X})\right]\right) - f(\boldsymbol{X})ln\left(\mathbb{E}_{X_i}\left[\mathbb{E}^i\left[f(\boldsymbol{X})\right]\right]\right)\right]\right] \leq \qquad (eq.(4.30)) \\
&\leq \sum_{i=1}^{n}\mathbb{E}\left[\mathcal{E}_{X_i}\left[f(\boldsymbol{X})\right]\right] = \mathbb{E}\left[\sum_{i=1}^{n}\mathcal{E}_{X_i}\left[f(\boldsymbol{X})\right]\right].
\end{aligned}$$

Notice that the main result extends to nonnegative random variables using the continuity of ϕ (the function introduced in the definition of entropy) at 0 and an elementary approximation argument. □

In order to prove the concentration of measure argument we need the logarithmic Sobolev inequality for Rademacher vectors.

Theorem 4.8.1 (Log-Sobolev inequality for Rademacher vectors)**.** *Let $f : \{-1,1\}^n \to \mathbb{R}$ be a function and $\boldsymbol{\epsilon}$ be an n-dimensional Rademacher vector* [1]*. Then*

$$\mathcal{E}(f^2(\boldsymbol{\epsilon})) \leq \frac{1}{2}\mathbb{E}\left[\sum_{i=1}^{n}\left(f(\boldsymbol{\epsilon}) - f(\overline{\boldsymbol{\epsilon}}^{(i)})\right)^2\right],$$

where $\overline{\boldsymbol{\epsilon}}^{(i)} = (\epsilon_1, \ldots, \epsilon_{i-1}, -\epsilon_i, \epsilon_{i+1}, \ldots, \epsilon_n)$.

Proof. First, we are going to show that

$$\mathcal{E}_{\epsilon_i}\left[f^2(\boldsymbol{\epsilon})\right] \leq \frac{1}{2}\mathbb{E}_{\epsilon_i}\left[\left(f(\boldsymbol{\epsilon}) - f(\overline{\boldsymbol{\epsilon}}^{(i)})\right)^2\right], \forall i \in \{1, \ldots, n\}. \tag{4.32}$$

Suppose that we have an arbitrary realization of the vector $(\epsilon_1, \ldots, \epsilon_{i-1}, \epsilon_{i+1}, \ldots, \epsilon_n)$. Since ϵ_i is a Rademacher random variable, the vector $(\epsilon_1, \ldots, \epsilon_{i-1}, \epsilon_i, \epsilon_{i+1}, \ldots, \epsilon_n)$

[1] A Rademacher random variable ϵ_i takes the values +1,-1, with probability 1/2 for each event. A Rademacher random vector $\boldsymbol{\epsilon}$ is a vector of independent Rademacher variables.

and as a result the function $f(\boldsymbol{\epsilon})$ can take two possible values, which we denote as $a, b \in \mathbb{R}$. Notice that $f(\overline{\boldsymbol{\epsilon}}^{(i)})$ can take the same two possible values.

Using the definition of conditional entropy we can deduce that,

$$\begin{aligned}\mathcal{E}_{\epsilon_i}\left[f^2(\boldsymbol{\epsilon})\right] &= \mathbb{E}_{\epsilon_i}\left[f^2(\boldsymbol{\epsilon}) ln(f^2(\boldsymbol{\epsilon}))\right] - \mathbb{E}_{\epsilon_i}\left[f^2(\boldsymbol{\epsilon})\right] ln\left(\mathbb{E}_{\epsilon_i}\left[f^2(\boldsymbol{\epsilon})\right]\right) \\ &= \frac{1}{2}\left(a^2 ln(a^2) + b^2 ln(b^2)\right) - \frac{a^2+b^2}{2} ln\left(\frac{a^2+b^2}{2}\right).\end{aligned}$$

Also,

$$\frac{1}{2}\mathbb{E}_{\epsilon_i}\left[\left(f(\boldsymbol{\epsilon}) - f(\overline{\boldsymbol{\epsilon}}^{(i)})\right)^2\right] = \frac{1}{2}\left(\frac{1}{2}(a-b)^2 + \frac{1}{2}(b-a)^2\right) = \frac{1}{2}(a-b)^2.$$

Thus, we want to show that

$$\frac{1}{2}\left(a^2 ln(a^2) + b^2 ln(b^2)\right) - \frac{a^2+b^2}{2} ln\left(\frac{a^2+b^2}{2}\right) \leq \frac{1}{2}(a-b)^2.$$

We define the function

$$g(x) = \frac{1}{2}\left(x^2 ln(a^2) + b^2 ln(b^2)\right) - \frac{x^2+b^2}{2} ln\left(\frac{x^2+b^2}{2}\right) - \frac{1}{2}(x-b)^2.$$

We calculate the first and second derivatives of $g(a)$

$$g'(x) = x ln\left(\frac{2x^2}{x^2+b^2}\right) - (x-b),$$

$$g''(x) = 1 + ln\left(\frac{2x^2}{x^2+b^2}\right) - \frac{2x^2}{x^2+b^2}.$$

We notice that $g(b) = g'(b) = 0$ and $g''(x) \leq 0, \forall\, x \in \mathbb{R}$. The last expression was obtained using the inequality $ln(x) \leq x - 1$. So we have established that $g(b) = 0$ is a global maximum and consequently $g(x) \leq 0$, $\forall x \in \mathbb{R}$. Therefore, equation (4.32) is valid for all $i \in \{1, \ldots, n\}$.

From proposition 4.8.1 we can obtain the following result

$$\mathcal{E}(f^2(\boldsymbol{\epsilon})) \leq \mathbb{E}\left[\sum_{i=1}^{n} \mathcal{E}_{\epsilon_i}\left[f^2(\boldsymbol{\epsilon})\right]\right]. \tag{4.33}$$

Combining proposition 4.8.1 and equation (4.33) yields the result provided in the definition.

□

Theorem 4.8.2 (Gaussian logarithmic Sobolev inequality)**.** *Let $f : \mathbb{R}^n \to \mathbb{R}$ be a continuously differentiable function such that $\mathbb{E}\left[\phi(f^2(\boldsymbol{g}))\right] < \infty$, where $\boldsymbol{g} \in \mathbb{R}^n$ is a standard Gaussian random vector. Then it holds that*

$$\mathcal{E}(f^2(\boldsymbol{g})) \leq 2\mathbb{E}\left[\|\nabla f(\boldsymbol{g})\|_2^2\right]$$

Proof. We start by proving the above result for $n = 1$ and $\boldsymbol{g} = g$, where g is a standard Gaussian random variable. Suppose that f has compact support. For the modulus of continuity $\omega(f', \delta) = \sup_{|t-u| \leq \delta} |f'(t) - f'(u)|$ we have that $\lim_{\delta \to 0} \omega(f', \delta) = 0$, because f' is uniformly continuous.

Let define

$$S_m = \frac{1}{\sqrt{m}} \sum_{i=1}^{m} \epsilon_i,$$

where $\boldsymbol{\epsilon} = (\epsilon_1, \ldots, \epsilon_m)$ is a Rademacher vector. Also, consider the function $f_1(\boldsymbol{\epsilon}) = f(S_m)$. Using theorem 4.8.1 we can attain

$$\begin{aligned} \mathcal{E}(f^2(S_m)) &\leq \frac{1}{2}\mathbb{E}\left[\sum_{i=1}^{m} \left(f_1(\boldsymbol{\epsilon}) - f_1(\overline{\boldsymbol{\epsilon}}^{(i)})\right)^2\right] = \\ &= \frac{1}{2}\mathbb{E}\left[\sum_{i=1}^{m} \left(f(S_m) - f\left(S_m - \frac{2\epsilon_i}{\sqrt{m}}\right)\right)^2\right], \end{aligned}$$

where $\overline{\boldsymbol{\epsilon}}^{(i)} = (\epsilon_1, \ldots, \epsilon_{i-1}, -\epsilon_i, \epsilon_{i+1}, \ldots, \epsilon_m)$.

Isolating and working with an arbitrary term of the above sum yields

$$\begin{aligned} \left|f(S_m) - f\left(S_m - \frac{2\epsilon_i}{\sqrt{m}}\right)\right| &= \left|\frac{2\epsilon_i}{\sqrt{m}} f'(S_m) + \int_{S_m - 2\epsilon_i/\sqrt{m}}^{S_m} (f'(t) - f'(S_m))\, dt\right| \leq \\ &\leq \frac{2}{\sqrt{m}} |f'(S_m)| + \frac{2}{\sqrt{m}} \omega(f', \frac{2}{\sqrt{m}}). \end{aligned}$$

Squaring both parts of the previous expression leads to

$$\left(f(S_m) - f\left(S_m - \frac{2\epsilon_i}{\sqrt{m}}\right)\right)^2 \leq \frac{4}{m} f'(S_m)^2 + \frac{8}{m} |f'(S_m)|\, \omega(f', \frac{2}{\sqrt{m}}) + \frac{4}{m} \omega(f', \frac{2}{\sqrt{m}})^2.$$

Therefore, we have that

$$\sum_{i=1}^{m} \left(f(S_m) - f\left(S_m - \frac{2\epsilon_i}{\sqrt{m}}\right)\right)^2 \leq 4 f'(S_m)^2 + 8 |f'(S_m)|\, \omega(f', \frac{2}{\sqrt{m}}) + 4\omega(f', \frac{2}{\sqrt{m}})^2.$$

Using the central limit theorem 4.2.2 and taking into account that f and f' are bounded function we can deduce that

$$\lim_{m \to \infty} \mathbb{E}\left[f'(S_m)^2\right] = \mathbb{E}\left[f'(g)^2\right]$$

and

$$\lim_{m \to \infty} \mathcal{E}(f^2(S_m)) = \mathcal{E}(f^2(g)),$$

where g is a standard Gaussian random variable.

Therefore,

$$\lim_{m\to\infty}\mathcal{E}(f^2(S_m)) \le \lim_{m\to\infty}\frac{1}{2}\mathbb{E}\left[4f'(S_m)^2 + 8\left|f'(S_m)\right|\omega(f',\frac{2}{\sqrt{m}}) + 4\omega(f',\frac{2}{\sqrt{m}})^2\right]$$
$$\Rightarrow \mathcal{E}(f^2(g)) \le 2\mathbb{E}\left[f'(g)^2\right].$$

Moving to the general case, assume that f does not necessarily have compact support. The expression $\mathbb{E}\left[\phi(f^2(g))\right] \le \infty$ implies that for a given $\epsilon > 0$ there exists $T > 0$, such that for any subset $I \subseteq \mathbb{R}\setminus[-T,T]$,

$$\frac{1}{\sqrt{2\pi}}\int_I \left|\phi(f(t)^2)\right| e^{-t^2/2}dt \le \epsilon \tag{4.34}$$

and

$$\frac{1}{\sqrt{2\pi}}\int_I e^{-t^2/2}dt \le \epsilon. \tag{4.35}$$

Also, let h be a continuously differentiable function, such that $0 \le h(t) \le 1$ and

$$h(t) = \begin{cases} 1 & , t \in [-T,T] \\ 0 & , t \notin [-T,T] \end{cases}.$$

Then, we define the function $\hat{f} = fh$. Notice that $\hat{f}$ is a continuously differentiable function with compact support. Consequently, using the results we proved in the first part of this proof,

$$\mathcal{E}(\hat{f}(g)^2) \le 2\mathbb{E}\left[\hat{f}'(g)^2\right].$$

Using the fact that entropy is a subadditive function (lemma 4.8.1(c)) we can see that

$$\begin{aligned}\mathcal{E}(f(g)^2) &= \mathcal{E}\left(\hat{f}(g)^2 + f(g)^2\left(1-h(g)^2\right)\right) \le \\ &\le \mathcal{E}\left(\hat{f}(g)^2\right) + \mathcal{E}\left(f(g)^2\left(1-h(g)^2\right)\right).\end{aligned} \tag{4.36}$$

In order to obtain a bound on the second term of the above expression we consider the sets $I_1 = \left\{t \in \mathbb{R}\setminus[-T,T] : f(t)^2 < e\right\}$ and $I_2 = \left\{t \in \mathbb{R}\setminus[-T,T] : f(t)^2 \ge e\right\}$. Then

$$\begin{aligned}\mathbb{E}\left[f(g)^2(1-h(g)^2)\right] &= \frac{1}{\sqrt{2\pi}}\int_{\mathbb{R}\setminus[-T,T]} f(g)^2(1-h(g)^2)e^{-t^2/2}dt \le \\ &\le \frac{1}{\sqrt{2\pi}}\int_{I_1} f(g)^2 e^{-t^2/2}dt + \frac{1}{\sqrt{2\pi}}\int_{I_2} f(g)^2 e^{-t^2/2}dt \le \\ &\le \frac{1}{\sqrt{2\pi}}\int_{I_1} e\cdot e^{-t^2/2}dt + \frac{1}{\sqrt{2\pi}}\int_{I_2}\left|\phi(f(g)^2)\right| e^{-t^2/2}dt \le \quad (eq.(4.34),(4.35)) \\ &\le (e+1)\epsilon.\end{aligned} \tag{4.37}$$

As a result,

$$\left|\phi\left(\mathbb{E}\left[f(g)^2(1-h(g)^2)\right]\right)\right| \leq \phi\left((e+1)\epsilon\right),$$

for sufficiently small ϵ.

Also, we introduce the sets $I_3 = \left\{t \in \mathbb{R} \setminus [-T,T] : f(t)^2(1-h(g)^2) < e\right\}$, $I_4 = \left\{t \in \mathbb{R} \setminus [-T,T] : f(t)^2(1-h(g)^2) \geq e\right\}$ and the value $\phi_0 = \max_{0 \leq t \leq e} |\phi(t)| = e^{-1}$.

Then,

$$\begin{aligned}
\left|\mathbb{E}\left[\phi\left(f(g)^2(1-h(g)^2)\right)\right]\right| &= \frac{1}{\sqrt{2\pi}}\left|\int\limits_{\mathbb{R}\setminus[-T,T]} \phi(f(g)^2(1-h(g)^2))e^{-t^2/2}dt\right| \leq \\
&\leq \frac{1}{\sqrt{2\pi}}\int\limits_{I_3} \phi_0 \cdot e^{-t^2/2}dt + \frac{1}{\sqrt{2\pi}}\int\limits_{I_4} \left|\phi(f(g)^2)\right| e^{-t^2/2}dt \leq \\
&\leq \phi_0\epsilon + \int\limits_{I_4} \left|\phi(f(g)^2)\right| e^{-t^2/2}dt \leq \qquad \text{(eq.(4.34), (4.35))} \\
&\leq (\phi_0+1)\epsilon.
\end{aligned}$$

Therefore, using the definition of entropy (definition 4.8.1) we can conclude that

$$\begin{aligned}
\mathcal{E}\left(f(g)^2\left(1-h(g)^2\right)\right) &\leq \left|\mathbb{E}\left[\phi\left(f(g)^2(1-h(g)^2)\right)\right]\right| + \left|\phi\left(\mathbb{E}\left[f(g)^2(1-h(g)^2)\right]\right)\right| \leq \\
&\leq |\phi((e+1)\epsilon)| + (\phi_0+1)\epsilon.
\end{aligned}$$

So, we have obtained a bound on the 2nd term of (4.36). Next we are going to improve the bound on the first term of expression (4.36). We have already shown that

$$\mathcal{E}(\hat{f}(g)^2) \leq 2\mathbb{E}\left[\hat{f}'(g)^2\right].$$

We can use the triangle inequality to obtain that

$$\sqrt{\mathbb{E}\left[\hat{f}'(g)^2\right]} = \sqrt{\mathbb{E}\left[(f'h + fh')(g)^2\right]} \leq \sqrt{\mathbb{E}\left[(f'h)(g)^2\right]} + \sqrt{\mathbb{E}\left[(fh')(g)^2\right]}.$$

For the first term of the above expression we can easily see that

$$\mathbb{E}\left[(f'h)(g)^2\right] = \mathbb{E}\left[f'(g)^2h(g)^2\right] \leq \mathbb{E}\left[f'(g)^2\right].$$

Moving to the second term we can show that

$$\begin{aligned}
\mathbb{E}\left[(fh')(g)^2\right] &= \frac{1}{\sqrt{2\pi}}\int\limits_{-\infty}^{+\infty} f(t)^2h'(t)^2e^{-t^2/2}dt \leq \frac{\|h'\|_\infty^2}{\sqrt{2\pi}}\int\limits_{I_1\cup I_2} f(t)^2e^{-t^2/2}dt \leq \\
&\leq \|h'\|_\infty^2(e+1)\epsilon.
\end{aligned}$$

In the last inequality we used the same reasoning as in expression (4.37). Therefore, we can obtain the following bound

$$\mathcal{E}(\hat{f}(g)^2) \leq 2\mathbb{E}\left[\hat{f}'(g)^2\right] = 2\sqrt{\mathbb{E}\left[\hat{f}'(g)^2\right]}^2 \leq$$
$$\leq 2\left(\sqrt{\mathbb{E}\left[f'(g)^2\right]} + \|h'\|_\infty \sqrt{(e+1)\epsilon}\right)^2.$$

For the entropy of f^2 we can obtain the following bound combining several of the previous results

$$\mathcal{E}(f(g)^2) \leq 2\left(\sqrt{\mathbb{E}\left[f'(g)^2\right]} + \|h'\|_\infty \sqrt{(e+1)\epsilon}\right)^2 + |\phi((e+1)\epsilon)| + (\phi_0 + 1)\epsilon.$$

The above result result holds for all $\epsilon > 0$. Combining that with the fact that $\lim_{t\to 0} \phi(t) = \phi(0) = 0$ we can conclude that

$$\mathcal{E}(f(g)^2) \leq 2\mathbb{E}\left[f'(g)^2\right].$$

Finally, we can generalize the previous result for any $n \in \mathbb{N}$ using the tensorization inequality (proposition 4.8.1),

$$\mathcal{E}(f^2(\boldsymbol{g})) \leq \mathbb{E}\left[\sum_{i=1}^n \mathcal{E}_{g_i}\left[f(\boldsymbol{g})\right]\right] \leq 2\mathbb{E}\left[\sum_{i=1}^n \left(\frac{\vartheta f}{\vartheta x_i}(\boldsymbol{g})\right)^2\right] =$$
$$= 2\mathbb{E}\left[\|\nabla f(\boldsymbol{g})\|_2^2\right].$$

□

Finally, we are ready to state and prove a concentration of measure argument.

Theorem 4.8.3 (Concentration of measure)**.** *Let $f : \mathbb{R}^n \to \mathbb{R}$ be a Lipschitz function (with constant L) and $\boldsymbol{g} \in \mathbb{R}^n$ be a standard Gaussian random vector. Then,*

$$\mathbb{P}(f(\boldsymbol{g}) - \mathbb{E}\left[f(\boldsymbol{g})\right] \geq t) \leq exp\left(-\frac{t^2}{2L^2}\right), \forall t > 0$$

and

$$\mathbb{P}(|f(\boldsymbol{g}) - \mathbb{E}\left[f(\boldsymbol{g})\right]| \geq t) \leq 2\, exp\left(-\frac{t^2}{2L^2}\right), \forall t > 0.$$

Proof. Assume that f is differentiable and let $t > 0$. Then, using the fact that f is a Lipschitz function we can see that $\|\nabla f(\boldsymbol{x})\|_2 \leq L$, $\forall\, \boldsymbol{x} \in \mathbb{R}^n$. Also, let define the function $h(\boldsymbol{x}) = e^{tf(\boldsymbol{x})/2}$, $\boldsymbol{x} \in \mathbb{R}^n$. Using the same fact we can easily deduce that $e^{tf(\boldsymbol{x})} \leq e^{t|f(\boldsymbol{0})|} e^{Lt\|\boldsymbol{x}\|_2}$, $\forall \boldsymbol{x} \in \mathbb{R}^n$ and thus $h(\boldsymbol{x})$ satisfies condition $\mathbb{E}\left[\phi(h^2(\boldsymbol{g}))\right] = \mathbb{E}\left[\phi(e^{tf(\boldsymbol{g})})\right] < \infty$. Therefore, h satisfies the conditions of theorem 4.8.2. Thus, using the logarithmic Sobolev inequality (theorem 4.8.2) on the function $h(\boldsymbol{g})$ yields

$$\mathcal{E}(e^{tf(\boldsymbol{g})}) \leq 2\mathbb{E}\left[\|\nabla e^{tf(\boldsymbol{g})/2}\|_2^2\right] = 2\mathbb{E}\left[\|\frac{t}{2}e^{tf(\boldsymbol{g})/2}\nabla f(\boldsymbol{g})\|_2^2\right] =$$
$$= \frac{t^2}{2}\mathbb{E}\left[e^{tf(\boldsymbol{g})}\|\nabla f(\boldsymbol{g})\|_2^2\right] \leq \frac{t^2L^2}{2}\mathbb{E}\left[e^{tf(\boldsymbol{g})}\right].$$

So, we have found an inequality of the form

$$\mathcal{E}\left(e^{tX}\right) \leq g(t)\mathbb{E}\left[e^{tX}\right],$$

with $g(t) = \frac{L^2t^2}{2}$ and $X = f(\boldsymbol{g})$. As a result, we can obtain the following bound for $t > 0$,

$$\mathbb{P}\left[f(\boldsymbol{g}) - \mathbb{E}\left[f(\boldsymbol{g})\right] \geq t\right] \leq \frac{exp\left(s\int\limits_0^s z^{-2}g(z)dz\right)}{e^{st}}$$
$$\Rightarrow \mathbb{P}\left[f(\boldsymbol{g}) - \mathbb{E}\left[f(\boldsymbol{g})\right] \geq t\right] \leq e^{s^2L^2/2-st}$$
$$\Rightarrow \mathbb{P}\left[f(\boldsymbol{g}) - \mathbb{E}\left[f(\boldsymbol{g})\right] \geq t\right] \leq \inf_{s>0}\left\{e^{s^2L^2/2-st}\right\}$$
$$\Rightarrow \mathbb{P}\left[f(\boldsymbol{g}) - \mathbb{E}\left[f(\boldsymbol{g})\right] \geq t\right] \leq exp\left(-\frac{t^2}{2L^2}\right). \qquad (s = \frac{t}{L^2} \text{ global minimum})$$

If we replace f with $-f$ we obtain the bound

$$\mathbb{P}\left[\mathbb{E}\left[f(\boldsymbol{g})\right] - f(\boldsymbol{g}) \geq t\right] \leq exp\left(-\frac{t^2}{2L^2}\right).$$

Therefore, the union bound establishes that

$$\mathbb{P}(|f(\boldsymbol{g}) - \mathbb{E}\left[f(\boldsymbol{g})\right]| \geq t) \leq 2exp\left(-\frac{t^2}{2L^2}\right), \forall t > 0.$$

Now moving to the case where f is not differentiable, for each $\epsilon > 0$ we can find a differentiable function $h(\boldsymbol{x})$, which is Lipschitz with the same constant L as f, such that

$$|f(\boldsymbol{x}) - h(\boldsymbol{x})| \leq \epsilon, \forall\, \boldsymbol{x} \in \mathbb{R}^n.$$

Consequently,

$$\mathbb{P}\left[f(\boldsymbol{g}) - \mathbb{E}\left[f(\boldsymbol{g})\right] \geq t\right] \leq \mathbb{P}\left[h(\boldsymbol{g}) - \mathbb{E}\left[h(\boldsymbol{g})\right] \geq t - 2\epsilon\right] \leq exp\left(-\frac{(t-2\epsilon)^2}{2L^2}\right)$$

The previous expression holds for every $\epsilon > 0$, hence we can establish that the basic result is also valid for non-differentiable functions. □

Chapter 5

Sparse vectors recovery with random matrices

5.1 Introduction

In chapter 2 we introduced the problem of sparse vector recovery and its extension to noisy scenarios. In both settings we highlighted the importance of choosing correctly the measurement matrix in order to ensure that the algorithm we use succesfully recovers all k-sparse vectors. In this chapter we are going to present in a mathematically rigorous way some results that refer to the ability of specific classes of random matrices to guarantee succesfull recovery of sparse vectors, under certain conditions, using the l_1 minimization task. This whole section is mainly based on [FR13], essentially providing a partial presentation of chapter 9.

There are two types of recovery results, uniform and non-uniform recovery guarantees. In *uniform recovery guarantees* the algorithm is able to recover successfully all k-sparse vector (after projection using a randomly drawn measurement matrix A), using the same matrix A, with high probability on the choice of matrix A. A uniform recovery guarantee roughly contains a statement of the form [FR13],

$$\mathbb{P}\left(\forall\, \boldsymbol{x} \in \Sigma_k\,,\ \text{ recovery of } \boldsymbol{x} \text{ is sucessfull using matrix } A\right) \geq 1-\epsilon.$$

On the other hand, in *non-uniform recovery guarantees* the algorithm is able to recover successfully a fixed k-sparse vector (after projection using a randomly drawn measurement matrix A), with high probability on the choice of matrix A. A non-uniform recovery guarantee contains a statement of the form [FR13],

$$\forall \boldsymbol{x} \in \Sigma_k \ :\ \mathbb{P}\left(\text{ recovery of } \boldsymbol{x} \text{ is sucessfull using matrix } A\right) \geq 1-\epsilon.$$

In the present chapter we are going to prove the following results:

- A uniform recovery result for subgaussian matrices using the l_1 minimization task (section 5.2).
- A non-uniform recovery result for Gaussian matrices using the l_1 minimization task (section 5.3).

5.2 Uniform recovery with subgaussian matrices

In this section we are going to prove a uniform recovery result for subgaussian matrices using the l_1 minimization task. First of all, we are going to show that if a random matrix A satisfies concentration inequality (5.1) then for that matrix the expression (5.2) holds, which constitutes an intermediate result in order to prove the R.I.P for A.

Theorem 5.2.1. *Let $A \in \mathbb{R}^{m\times n}$ be a random matrix for which the following concentration inequality holds*

$$\mathbb{P}\left(\left|\|A\boldsymbol{x}\|_2^2 - \|\boldsymbol{x}\|_2^2\right| \geq t\|\boldsymbol{x}\|_2^2\right) \leq 2e^{-c_0 m t^2},\ \forall \boldsymbol{x} \in \mathbb{R}^n,\ t \in (0,1). \tag{5.1}$$

Given $\delta, \epsilon \in (0,1)$ and $S \subseteq \{1,\dots,n\}$ with $card(S) = k$, if

$$m \geq \frac{14k + 4\ln(2\epsilon^{-1})}{3c_0\delta^2}$$

then

$$\|A_S^T A_S - I\|_{2\to 2} < \delta, \tag{5.2}$$

[1] *with probability at least $1-\epsilon$.*

Proof. We perform a discretization of the unit ball of vectors with at most k non-zero elements in the positions specified by the index set S. Specifically, we consider the set $\mathcal{B}_S = \{\boldsymbol{x} \in \mathbb{R}^n : \|\boldsymbol{x}\|_2 \leq 1 \wedge supp(\boldsymbol{x}) \subseteq S\}$. From theorem (A.4.1) about covering numbers we know that, given a radius $\rho \in (0, 1/2)$, we can find a finite set $U \subseteq \mathcal{B}_S$, such that the following expressions hold

$$card(U) \leq \left(1 + \frac{2}{\rho}\right)^k \tag{5.3}$$

and

$$\min_{\boldsymbol{u}\in U} \|\boldsymbol{z} - \boldsymbol{u}\|_2 \leq \rho,\ \forall\, \boldsymbol{z} \in \mathcal{B}_S.$$

Using concentration inequality (5.1) it is easy to see that

$$\begin{aligned}
\mathbb{P}\left(\left|\|A\boldsymbol{u}\|_2^2 - \|\boldsymbol{u}\|_2^2\right| \geq t\|\boldsymbol{u}\|_2^2 \text{ for some } u \in U\right) &\leq \sum_{\boldsymbol{u}\in U} \mathbb{P}\left(\left|\|A\boldsymbol{u}\|_2^2 - \|\boldsymbol{u}\|_2^2\right| \geq t\|\boldsymbol{u}\|_2^2\right) \\
&\leq 2\, card(U) e^{-c_0 m t^2} \\
&\leq 2\left(1 + \frac{2}{\rho}\right)^k e^{-c_0 m t^2}. \qquad \text{(eq. (5.3))}
\end{aligned}$$

Suppose that for a randomly drawn random matrix A, from the probability distribution for which the inequality (5.1) holds, we have that

$$\left|\|A\boldsymbol{u}\|_2^2 - \|\boldsymbol{u}\|_2^2\right| < t\|\boldsymbol{u}\|_2^2, \forall \boldsymbol{u} \in U.$$

Also, we set $B = A_S^T A_S - I$. Then, we can write the previous expression as

$$|\langle B\boldsymbol{u}, \boldsymbol{u}\rangle| < t, \forall \boldsymbol{u} \in U. \tag{5.4}$$

[1] A_S denotes the matrix created by keeping the columns of matrix A specified by the index set S and removing the rest.

The previous expression holds with the following probability,

$$\mathbb{P}\left(\left|\|A\boldsymbol{u}\|_2^2 - \|\boldsymbol{u}\|_2^2\right| < t\|\boldsymbol{u}\|_2^2, \forall \boldsymbol{u} \in U\right) \geq 1 - 2\left(1 + \frac{2}{\rho}\right)^k e^{-c_0 m t^2}. \tag{5.5}$$

We want to show that for appropriate choice of ρ and t expression (5.4) implies

$$\left|\|A\boldsymbol{x}\|_2^2 - \|\boldsymbol{x}\|_2^2\right| < \delta\|\boldsymbol{x}\|_2^2 \leq \delta, \ \forall \boldsymbol{x} \in \mathcal{B}_S$$

which in turn implies

$$|\langle B\boldsymbol{x}, \boldsymbol{x}\rangle| < \delta, \forall \boldsymbol{x} \in \mathcal{B}_S$$

and

$$\|B\|_{2\to 2} < \delta.$$

So, let $\boldsymbol{x} \in \mathcal{B}_S$ and $\boldsymbol{u} \in U$, such that $\|\boldsymbol{x} - \boldsymbol{u}\|_2 \leq \rho < 1/2$. Then we have that

$$\begin{aligned}
|\langle B\boldsymbol{x}, \boldsymbol{x}\rangle| &= |\langle B\boldsymbol{x}, \boldsymbol{x}\rangle + \langle B\boldsymbol{u}, \boldsymbol{u}\rangle - \langle B\boldsymbol{u}, \boldsymbol{u}\rangle + \langle \boldsymbol{x}, B\boldsymbol{u}\rangle - \langle \boldsymbol{x}, B\boldsymbol{u}\rangle| = \\
&= |\langle B\boldsymbol{x}, \boldsymbol{x}\rangle + \langle B\boldsymbol{u}, \boldsymbol{u}\rangle - \langle B\boldsymbol{u}, \boldsymbol{u}\rangle + \langle B\boldsymbol{u}, \boldsymbol{x}\rangle - \langle B\boldsymbol{x}, \boldsymbol{u}\rangle| = && (B \text{ symmetric}),(\text{commutativity}) \\
&= |\langle B\boldsymbol{u}, \boldsymbol{u}\rangle + \langle B(\boldsymbol{x} + \boldsymbol{u}), \boldsymbol{x} - \boldsymbol{u}\rangle| \leq \\
&\leq |\langle B\boldsymbol{u}, \boldsymbol{u}\rangle| + |\langle B(\boldsymbol{x} + \boldsymbol{u}), \boldsymbol{x} - \boldsymbol{u}\rangle| \leq \\
&\leq t + \|B(\boldsymbol{x} + \boldsymbol{u})\|_2\|\boldsymbol{x} - \boldsymbol{u}\|_2 < && (\text{Cauchy-Schwarz ineq.}), (\text{eq. } 5.4) \\
&< t + \|B\|_{2\to 2}\|\boldsymbol{x} + \boldsymbol{u}\|_2\|\boldsymbol{x} - \boldsymbol{u}\|_2 \leq \\
&\leq t + \|B\|_{2\to 2}(\|\boldsymbol{x}\|_2 + \|\boldsymbol{u}\|_2)\rho \leq \\
&\leq t + 2\rho\|B\|_{2\to 2}.
\end{aligned}$$

We know that we can obtain the norm of B as

$$\|B\|_{2\to 2} = \sup_{\boldsymbol{x} \in \mathcal{B}_S} \langle B\boldsymbol{x}, \boldsymbol{x}\rangle .$$

Therefore, from the previous inequality we can conclude that

$$\|B\|_{2\to 2} < t + 2\|B\|_{2\to 2}\rho \Rightarrow \|B\|_{2\to 2} < \frac{t}{1 - 2\rho}, t \in (0, 1), \rho \in \left(0, \frac{1}{2}\right).$$

We choose $t = (1 - 2\rho)\delta$ and as a result we establish that $\|B\|_{2\to 2} < \delta$. For the probability that this event occurs we can use (5.5) to deduce that

$$\mathbb{P}\left(\|B\|_{2\to 2} < \delta\right) \geq 1 - 2\left(1 + \frac{2}{\rho}\right)^k e^{-c_0 m(1-2\rho)^2\delta^2}.$$

It is straightforward that

$$\mathbb{P}\left(\|B\|_{2\to 2} \geq \delta\right) \leq 2\left(1 + \frac{2}{\rho}\right)^k e^{-c_0 m(1-2\rho)^2\delta^2}. \tag{5.6}$$

So, if we denote $\epsilon = \mathbb{P}\left(\|B\|_{2\to 2} \geq \delta\right)$ we have that

$$\|B\|_{2\to 2} \leq \delta$$

occurs with probability at least $1-\epsilon$ provided

$$\begin{aligned}
&\epsilon \leq 2\left(1+\frac{2}{\rho}\right)^k e^{-c_0 m(1-2\rho)^2\delta^2} \Rightarrow \\
&m \geq \frac{1}{c_0(1-2\rho)^2\delta^2}\left(ln\left(1+\frac{2}{\rho}\right)k + ln\left(2\epsilon^{-1}\right)\right).
\end{aligned}$$

We can obtain the bound provided in the definition if we consider $\rho = \dfrac{2}{e^{7/2}-1}$, which leads to

$$m \geq \frac{14k + 4\ln(2\epsilon^{-1})}{3c_0\delta^2}.$$

□

In the following remark an alternative description of the R.I.P is provided that is necessary for the next theorem.

Remark 5.2.1 (Alternative description of R.I.P)**.** *Starting from the definition of the R.I.P provided in expression (2.6) we can obtain the following alternative description*

$$\begin{aligned}
&(1-\delta)\|\boldsymbol{x}\|_2^2 \leq \|A\boldsymbol{x}\|_2^2 \leq (1+\delta)\|\boldsymbol{x}\|_2^2,\ \forall \boldsymbol{x} \in \Sigma_k \\
&\Rightarrow \left|\|A_S\boldsymbol{x}\|_2^2 - \|\boldsymbol{x}\|_2^2\right| \leq \delta\|\boldsymbol{x}\|_2^2,\ \forall\, S \subseteq \{1,2,\ldots,n\},\ card(S) \leq k,\ \forall\, \boldsymbol{x} \in \mathbb{R}^S \\
&\Rightarrow \frac{\langle A_S\boldsymbol{x}, A_S\boldsymbol{x}\rangle - \langle \boldsymbol{x},\boldsymbol{x}\rangle}{\|\boldsymbol{x}\|_2^2} \leq \delta,\ \forall\, S \subseteq \{1,2,\ldots,n\},\ card(S) \leq k,\ \forall\, \boldsymbol{x} \in \mathbb{R}^S \setminus \{\mathbf{0}\} \\
&\Rightarrow \frac{\left\langle (A_S^TA_S - I)\boldsymbol{x},\boldsymbol{x}\right\rangle}{\|\boldsymbol{x}\|_2^2} \leq \delta,\ \forall\, S \subseteq \{1,2,\ldots,n\},\ card(S) \leq k,\ \forall\, \boldsymbol{x} \in \mathbb{R}^S \setminus \{\mathbf{0}\} \\
&\Rightarrow \max_{\boldsymbol{x}\in\mathbb{R}^S\setminus\{\mathbf{0}\}} \frac{\left\langle (A_S^TA_S - I)\boldsymbol{x},\boldsymbol{x}\right\rangle}{\|\boldsymbol{x}\|_2^2} \leq \delta,\ \forall\, S \subseteq \{1,2,\ldots,n\},\ card(S) \leq k \qquad (A_S^TA_S - I\ Symm.) \\
&\Rightarrow \|A_S^TA_S - I\|_{2\to 2} \leq \delta,\ \forall\, S \subseteq \{1,2,\ldots,n\},\ card(S) \leq k \\
&\Rightarrow \max_{S\subseteq\{1,2,\ldots,n\},\, card(S)\leq k} \|A_S^TA_S - I\|_{2\to 2} \leq \delta.
\end{aligned}$$

Therefore, the R.I.P constant of order k is

$$\delta_k = \max_{S\subseteq\{1,2,\ldots,k\},\, card(S)\leq k} \|A_S^TA_S - I\|_{2\to 2}.$$

Now we can use the previous theorem to establish the R.I.P condition for a random matrix that satisfies concentration inequality (5.1).

Theorem 5.2.2. *Let $A \in \mathbb{R}^{m\times n}$ be a random matrix for which the following concentration inequality holds*

$$\mathbb{P}\left(\left|\|A\boldsymbol{x}\|_2^2 - \|\boldsymbol{x}\|_2^2\right| \geq t\|\boldsymbol{x}\|_2^2\right) \leq 2e^{-c_0 m t^2},\ \forall \boldsymbol{x} \in \mathbb{R}^n,\ t \in (0,1) \qquad (5.7)$$

Given $\delta, \epsilon \in (0,1)$, if

$$m \geq \frac{2k\left(9 + 2ln\left(\frac{n}{k}\right)\right) + 4\ln(2\epsilon^{-1})}{3c_0\delta^2},$$

then the restricted isometry constant of A satisfies $\delta_k < \delta$ with probability at least $1-\epsilon$.

Proof. Taking into account remark 5.2.1 we can easily see that

$$\delta_k = \sup_{S\subseteq[n], card(S)=k} \|A_S^T A_S - I\|_{2\to 2} = \sup_{S\subseteq[n], card(S)=k} \|B\|_{2\to 2}$$

Therefore, we have that

$$\begin{aligned}
\mathbb{P}(\delta_k \geq \delta) &\leq \sum_{S\subseteq[n], card(S)=k} \mathbb{P}(\|B\|_{2\to 2} \geq \delta) \leq && \text{(Union bound)} \\
&\leq 2\binom{n}{k}\left(1+\frac{2}{\rho}\right)^k e^{-c_0\delta^2(1-2\rho)^2 m} \leq && \text{(eq. (5.6))} \\
&\leq 2\left(\frac{en}{k}\right)^k \left(1+\frac{2}{\rho}\right)^k e^{-c_0\delta^2(1-2\rho)^2 m}. && \text{(lemma (A.5.1))}
\end{aligned}$$

Let set $\epsilon = \mathbb{P}(\|B\|_{2\to 2} \geq \delta)$ and choose $\rho = \frac{2}{e^{7/2}-1}$. Then we can see that $\delta_k < \delta$ holds with probability $1-\epsilon$ if

$$\begin{aligned}
&\epsilon \leq 2\left(\frac{en}{k}\right)^k \left(1+\frac{2}{\rho}\right)^k e^{-c_0\delta^2(1-2\rho)^2 m} \Rightarrow && \text{(lemma (A.5.1))} \\
&m \geq \frac{2k\left(9 + 2ln\left(\frac{n}{k}\right)\right) + 4\ln(2\epsilon^{-1})}{3c_0\delta^2}.
\end{aligned}$$

□

Next, we are intending to show that a matrix with independent, subgaussian, isotropic rows satisfy concentration inequality (5.7).

Proposition 5.2.1. *Let $A \in \mathbb{R}^{m\times n}$ be a random matrix with independent, isotropic and subgaussian rows (all rows have the same subgaussian parameter c). Then the following concentration inequality holds*

$$\mathbb{P}\left(\left|\frac{1}{m}\|A\boldsymbol{x}\|_2^2 - \|\boldsymbol{x}\|_2^2\right| \geq t\|\boldsymbol{x}\|_2^2\right) \leq 2e^{-c_0 m t^2}, \forall\, \boldsymbol{x} \in \mathbb{R}^n,\ t \in (0,1),$$

for a constant c_0 that depends only on c.

Proof. The matrix A is of the form

$$A = \begin{bmatrix} \boldsymbol{Y}_1^T \\ \boldsymbol{Y}_2^T \\ \vdots \\ \boldsymbol{Y}_m^T \end{bmatrix}, \tag{5.8}$$

where Y_i, $1 \leq i \leq m$ are independent, isotropic, subgaussian random vectors. We define the following finite sequence of random variables,

$$Z_i = |\langle \boldsymbol{Y}_i, \boldsymbol{x} \rangle|^2 - \|\boldsymbol{x}\|_2^2,\ 1 \leq i \leq m,$$

where $\boldsymbol{x} \in \mathbb{R}^n$ with $\|\boldsymbol{x}\|_2^2 = 1$ (we can assume that without loss of generality).

Using the fact that $\boldsymbol{Y}_i$ is an isotropic random vector (theorem 4.1.7) we can see that

$$\mathbb{E}\left[Z_i\right] = \mathbb{E}\left[|\langle \boldsymbol{Y}_i, \boldsymbol{x} \rangle|^2\right] - \mathbb{E}\left[\|\boldsymbol{x}\|_2^2\right] = \|\boldsymbol{x}\|_2^2 - \|\boldsymbol{x}\|_2^2 = 0,$$

for every $1 \leq i \leq m$.

Notice that the random variable $\langle \boldsymbol{Y}_i, \boldsymbol{x} \rangle$ is subgaussian, since Y_i is a subgaussian random vector. Hence, Z_i is also a subexponential random variable.

Also, we have that

$$\frac{1}{m}\|A\boldsymbol{x}\|_2^2 - \|\boldsymbol{x}\|_2^2 = \frac{1}{m}\sum_{i=1}^{m}\left(|\langle \boldsymbol{Y}_i, \boldsymbol{x} \rangle|^2 - \|\boldsymbol{x}\|_2^2\right) = \frac{1}{m}\sum_{i=1}^{m} Z_i.$$

The random vectors $\boldsymbol{Y}_i$ are independent and as a result the random variables Z_i are independent as well. We have established that Z_i satisfy all the conditions of Bernstein's inequality for subexponential random variables (theorem 4.4.1). Therefore, we can obtain the following result

$$\mathbb{P}\left(\left|\frac{1}{m}\sum_{i=1}^{m} Z_i\right| \geq t\right) = \mathbb{P}\left(\left|\sum_{i=1}^{m} Z_i\right| \geq mt\right) \leq 2exp\left(-\frac{\kappa^2 m^2 t^2/2}{2\beta m + \kappa m t}\right) =$$

$$= 2exp\left(-\frac{r^2}{4k + 2rt}mt^2\right) \leq 2exp\left(-\frac{r^2}{4k + 2r}mt^2\right). \quad (t \in [0,1])$$

Finally, we choose $c_0 = \dfrac{\kappa^2}{4\beta + 2\kappa t}$. As a result, we conclude

$$\mathbb{P}\left(\frac{1}{m}\|A\boldsymbol{x}\|_2^2 - \|\boldsymbol{x}\|_2^2 \geq t\|\boldsymbol{x}\|_2^2\right) \leq 2e^{-c_0 m t^2},\ \forall \boldsymbol{x} \in \mathbb{R}^n,\ t \in (0,1).$$

□

Using the previous results we can state rigorously the R.I.P condition for matrices with isotropic and independent subgaussian rows.

Theorem 5.2.3. *Let $A \in \mathbb{R}^{m \times n}$ be a random matrix with independent, isotropic, subgaussian rows (all rows have the same subgaussian parameter c). If*

$$m \geq \frac{C\left(k \ln\left(\frac{en}{k}\right) + \ln\left(2\epsilon^{-1}\right)\right)}{\delta^2},$$

then the matrix $\dfrac{1}{\sqrt{m}}A$ satisfies the R.I.P condition with a constant $\delta_k \leq \delta$ and with probability at least $1 - \epsilon$.

Proof. Combining proposition 5.2.1 with theorem 5.2.2 and possibly altering some constants establishes the above result. □

At this point we have established that a (normalized) matrix with isotropic, subgaussian rows under some specific conditions satisfy the R.I.P. What we need now is to show that this result also holds for subgaussian matrices. We are going to show that using the following lemma.

Lemma 5.2.1. *Suppose that $\boldsymbol{Y} \in \mathbb{R}^n$ is a random vector of independent subgaussian entries, with $\mathbb{E}[Y_i] = 0$, $Var[Y_i] = 1$, $1 \leq i \leq n$ and the same subgaussian parameter c. Then $\boldsymbol{Y}$ is an isotropic, subgaussian random vector with parameter c.*

Proof. From the definition we know that Y_i are independent subgaussian random variables with $\mathbb{E}[Y_i] = 0$ and $Var[Y_i] = 1$. Obviously,

$$\mathbb{E}[Y_i Y_j] = \begin{cases} 0 & i \neq j \\ 1 & i = j \end{cases}.$$

Then, for $\boldsymbol{x} \in \mathbb{R}^n$ we have that

$$\mathbb{E}\left[|\langle \boldsymbol{Y}, \boldsymbol{x}\rangle|^2\right] = \mathbb{E}\left[\left|\sum_{i=1}^n x_i Y_i\right|^2\right] = \mathbb{E}\left[\sum_{i=1}^n \sum_{j=1}^n x_i x_j Y_i Y_j\right] =$$
$$= \sum_{i=1}^n \sum_{j=1}^n x_i x_j \mathbb{E}[Y_i Y_j] = \sum_{i=1}^n x_i^2 = \|\boldsymbol{x}\|_2^2.$$

As a result, from definition (4.1.7) we deduce that $\boldsymbol{Y}$ is an isotropic random vector.

Also, considering $\boldsymbol{x} \in \mathbb{R}^n$, $\|\boldsymbol{x}\|_2 = 1$, proposition (4.3.3) establishes that the random variable

$$\langle \boldsymbol{Y}, \boldsymbol{x}\rangle = \sum_{i=1}^n x_i Y_i \tag{5.9}$$

is subgaussian with parameter $c\|\boldsymbol{x}\|_2^2 = c$ (independent of $\boldsymbol{x} \in \mathbb{R}^n$, $\|\boldsymbol{x}\|_2 = 1$). Then, from definition (4.3.2) we can conclude that $\boldsymbol{Y}$ is a subgaussian random vector with parameter c.

□

We can now use the previous lemma to establish the R.I.P for subgaussian matrices.

Theorem 5.2.4. *Let $A \in \mathbb{R}^{m \times n}$ be a subgaussian matrix. There exists $C > 0$ (depending only on the subgaussian parameters k and r) such that if*

$$m \geq \frac{C\left(k \ln\left(\frac{en}{k}\right) + \ln\left(2\epsilon^{-1}\right)\right)}{\delta^2},$$

then the matrix $\frac{1}{\sqrt{m}}A$ satisfies the R.I.P condition with a constant $\delta_s \leq \delta$ and with probability at least $1 - \epsilon$.

Proof. This theorem follows immediately from theorem 5.2.3 and lemma 5.2.1.

□

Finally, we can present a uniform recovery guarantee for subgaussian matrices using the l_1 minimization task, in the noisy case.

Theorem 5.2.5 (Uniform recovery guarantee for subgaussian matrices-noisy case)**.** *Let $A \in \mathbb{R}^{m\times n}$ be a subgaussian random matrix and $\epsilon \in (0,1)$. There exist $c_1, c_2 > 0$ (depending only on the subgaussian parameters k, r) and $d_1, d_2 > 0$ (universal constants) such that if*

$$m \geq c_1 k \ln\left(\frac{en}{k}\right) + c_2 \ln\left(2\epsilon^{-1}\right),$$

then for all $\boldsymbol{x} \in \mathbb{R}^n$ a solution of the robust l_1 minimization task (2.9.2), with $\boldsymbol{y} = A\boldsymbol{x} + \boldsymbol{e}$ and $\|\boldsymbol{e}\|_2 \leq \sqrt{m}\eta$ (for some $\eta > 0$), satisfy the following conditions

$$\|\boldsymbol{x} - \hat{\boldsymbol{x}}\|_1 \leq d_1 \sigma_k(\boldsymbol{x})_1 + d_2\sqrt{k}\eta$$

and

$$\|\boldsymbol{x} - \hat{\boldsymbol{x}}\|_2 \leq d_1 \frac{\sigma_k(\boldsymbol{x})_1}{\sqrt{k}} + d_2\eta.$$

Proof. The robust l_1 minimization task (2.9.2), with $\|\boldsymbol{e}\|_2 \leq \sqrt{m}\eta$, is equivalent to the following optimization task

$$\begin{aligned} \underset{\boldsymbol{x}\in\mathbb{R}^n}{\text{minimize}} \quad & \|\boldsymbol{x}\|_1 \\ \text{subject to} \quad & \|\frac{1}{\sqrt{m}}\boldsymbol{y} - \frac{1}{\sqrt{m}}A\boldsymbol{x}\|_2 \leq \eta. \end{aligned}.$$

Therefore, combining theorems 5.2.4 and 2.9.3 provides the result we want. □

In the noiseless case we have the following theorem.

Theorem 5.2.6 (Uniform recovery guarantee for subgaussian matrices-noiseless case)**.** *Let $A \in \mathbb{R}^{m\times n}$ be a subgaussian random matrix and $\epsilon \in (0,1)$. There exist $c_1, c_2 > 0$ (depending only on k, r) such that if*

$$m \geq c_1 s \ln\left(\frac{en}{k}\right) + c_2 \ln\left(2\epsilon^{-1}\right) \tag{5.10}$$

then the l_1 minimization task (2.6.3), with $\boldsymbol{y} = A\boldsymbol{x}$, recovers $\boldsymbol{x}$, $\forall \boldsymbol{x} \in \Sigma_k$ with probability at least $1 - \epsilon$

Proof. Combining theorems 5.2.4 (in the special case where $\sigma_k(\boldsymbol{x})_p = 0$ and $\epsilon = 0$), 2.9.3 and taking into account that exact recovery is not affected by the normalization of the matrix provides the result we want. □

5.3 Non-uniform recovery with Gaussian matrices

In order to fully develop the theoretical results of this section we need to define the tangent cone of the l_1 norm.

Definition 5.3.1 (Tangent cone of l_1 norm)**.** *The tangent cone of the l_1 norm at $\boldsymbol{x} \in \mathbb{R}^n$ is*

$$\mathcal{T}(\boldsymbol{x}) = cone\left\{\boldsymbol{z} - \boldsymbol{x} : \boldsymbol{z} \in \mathbb{R}^n \text{ and } \|\boldsymbol{z}\|_1 \leq \|\boldsymbol{x}\|_1\right\},$$

where cone defines the conic hull of a set.

Essentially, the tangent cone of the l_1 norm is the set of all descent directions from $\boldsymbol{x}$, i.e. the set of all direction from $\boldsymbol{x}$ such that the l_1 norm decreases. We can now state a necessary and sufficient condition for the successful recovery of a k-sparse vector with the l_1 minimization task, using the tangent cone of the l_1 norm.

Theorem 5.3.1 (Sufficient and necessary condition for successful recovery-noiseless case)**.** *Let $\boldsymbol{x} \in \mathbb{R}^n$ and $A \in \mathbb{R}^{m \times n}$. Then optimization task (2.6.3), with $\boldsymbol{y} = A\boldsymbol{x}$, returns $\boldsymbol{x}$ as a unique solution, if and only if*

$$Ker(A) \cap \mathcal{T}(\boldsymbol{x}) = \{\boldsymbol{0}\}.$$

Proof. Suppose that $Ker(A) \cap \mathcal{T}(\boldsymbol{x}) = \{\boldsymbol{0}\}$ holds. Also, let $\hat{\boldsymbol{x}}$ denote a solution of (2.6.3) with $\boldsymbol{y} = A\boldsymbol{x}$. Then, we have that $\|\hat{\boldsymbol{x}}\|_1 \leq \|\boldsymbol{x}\|_1$ and $\boldsymbol{y} = A\hat{\boldsymbol{x}}$. Let, $\boldsymbol{z} = \hat{\boldsymbol{x}} - \boldsymbol{x}$. From definition (5.3.1) we can see that, $\boldsymbol{z} \in \mathcal{T}(\boldsymbol{x})$. Furthermore, $A\boldsymbol{x} = A\hat{\boldsymbol{x}} \Rightarrow A(\hat{\boldsymbol{x}} - \boldsymbol{x}) = 0 \Rightarrow \boldsymbol{z} \in Ker(A)$. As a result, $Ker(A) \cap \mathcal{T}(\boldsymbol{x}) = \{\boldsymbol{0}\} \Rightarrow \hat{\boldsymbol{x}} = \boldsymbol{x}$.

Now, suppose that $\boldsymbol{x}$ is the unique solution of (2.6.3) with $\boldsymbol{y} = A\boldsymbol{x}$. Let $\boldsymbol{z} \in \mathcal{T}(\boldsymbol{x}) \setminus \{\boldsymbol{0}\}$. We can write this vector as $\boldsymbol{z} = \sum_i c_i(\boldsymbol{u}_i - \boldsymbol{x})$, where $c_i \geq 0$ and $\|\boldsymbol{u}_i\|_1 \leq \|\boldsymbol{x}\|_1$. Notice that $\boldsymbol{z} \neq \boldsymbol{0} \Rightarrow \sum_i c_i > 0$. Suppose now that $\boldsymbol{z} \in KerA$. Then,

$$A\left(\sum_i c_i(\boldsymbol{u}_i - \boldsymbol{x})\right) = \boldsymbol{0} \Rightarrow A\left(\sum_i \frac{c_i}{\sum_j c_j}(\boldsymbol{u}_i - \boldsymbol{x})\right) = \boldsymbol{0} \Rightarrow A\left(\sum_i c_i' \boldsymbol{u}_i\right) = A\boldsymbol{x},$$

where $c_i' = \dfrac{c_i}{\sum_j c_j}$. Furthermore, $\|\sum_i c_i' \boldsymbol{u}_i\|_1 \leq \sum_i c_i' \|\boldsymbol{u}_i\|_1 \leq \|\boldsymbol{x}\|_1$. However, we have made the assumption that $\boldsymbol{x}$ is a unique solution of (2.6.3) and thus

$$\sum_i c_i' \boldsymbol{u}_i = \boldsymbol{x} \Rightarrow \boldsymbol{z} = 0.$$

So, we reached a contradiction. Therefore,

$$\boldsymbol{z} \notin KerA \Rightarrow KerA \cap (\mathcal{T}(\boldsymbol{x}) \setminus \{\boldsymbol{0}\}) = \emptyset \Rightarrow KerA \cap \mathcal{T}(\boldsymbol{x}) = \{\boldsymbol{0}\}.$$

□

In order for the content of the previous theorem to become more clear we provide figures (5.1) and (5.2). So, let $\boldsymbol{x}_0 \in \mathbb{R}^n$ and $A \in \mathbb{R}^{m \times n}$. We have successful recovery if and only if the affine space

$$X = \{\boldsymbol{x} \in \mathbb{R}^n : \boldsymbol{y} = A\boldsymbol{x}, \text{ with } \boldsymbol{y} = A\boldsymbol{x}_0\} = \boldsymbol{x}_0 + Ker(A)$$

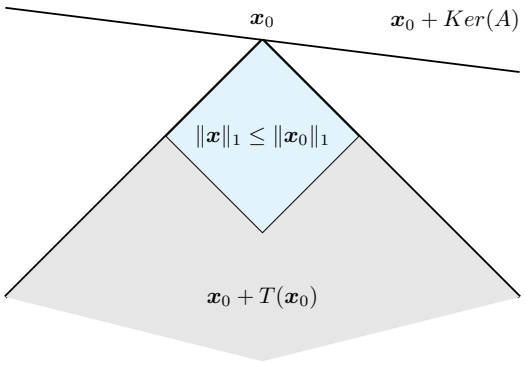

Figure 5.1: We have a sparse vector $\boldsymbol{x}_0 \in \mathbb{R}^n$, the affine space generated by the solutions of the equation $\boldsymbol{y} = A\boldsymbol{x}$, with $\boldsymbol{y} = A\boldsymbol{x}_0$, the l_1 norm ball (with radius $\|\boldsymbol{x}_0\|_1$) and the tangent cone of the l_1 norm at $\boldsymbol{x}_0$. Notice that we have successful recovery if and only if the affine space intersects the shifted tangent cone only at $\boldsymbol{x}_0$. In that case we can successfully recover $\boldsymbol{x}_0$ using the l_1 minimization task. The image is based on [Ame+14].

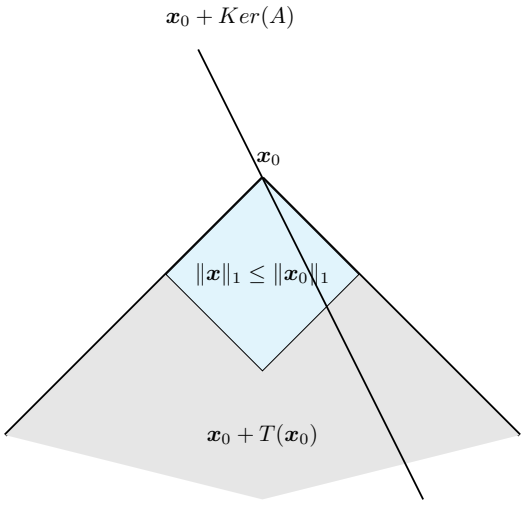

Figure 5.2: In this figure we share the same setting with figure (5.1). However, we illustrate the case where the l_1 minimization algorithm fails to recover $\boldsymbol{x}_0$.

intersects the (shifted) tangent cone $\mathcal{T}(\boldsymbol{x}_0)+\boldsymbol{x}_0$ only at $\{\boldsymbol{x}_0\}$. Equivalently, we have successful recovery if and only if

$$Ker(A)\cap\mathcal{T}(\boldsymbol{x}_0)=\{\boldsymbol{0}\}.$$

In the theoretical analysis that will follow, we are going to work with an alternative condition, equivalent with the initial one.

Proposition 5.3.1 (Alternative sufficient and necessary condition for successful recovery-noiseless case)**.** *Let $\boldsymbol{x}\in\mathbb{R}^n$ and $A\in\mathbb{R}^{m\times n}$. Then optimization task (2.6.3), with $\boldsymbol{y}=A\boldsymbol{x}$, returns $\boldsymbol{x}$ as a unique solution, if and only if*

$$\inf_{\boldsymbol{z}\in T(\boldsymbol{x})\cap S^{n-1}}\|A\boldsymbol{z}\|_2>0.$$

Proof. We are going to show that

$$Ker(A)\cap\mathcal{T}(\boldsymbol{x})=\{\boldsymbol{0}\}\Leftrightarrow\inf_{\boldsymbol{z}\in T(\boldsymbol{x})\cap S^{n-1}}\|A\boldsymbol{z}\|_2>0.$$

Assume that $Ker(A)\cap\mathcal{T}(\boldsymbol{x})=\{\boldsymbol{0}\}$ holds. Then, if $\boldsymbol{z}\in\mathcal{T}(\boldsymbol{x})\cap S^{n-1}$ it is obvious that $\boldsymbol{z}\notin Ker(A)$ and thus $\|A\boldsymbol{z}\|_2>0$. As a result,

$$\inf_{\boldsymbol{z}\in T(\boldsymbol{x})\cap S^{n-1}}\|A\boldsymbol{z}\|_2>0.$$

Now assume that $\inf_{\boldsymbol{z}\in\mathcal{T}(\boldsymbol{x})\cap S^{n-1}}\|A\boldsymbol{z}\|_2>0$. Remember that $T(\boldsymbol{x})$ is a cone, hence if $\boldsymbol{y}\in\mathcal{T}(\boldsymbol{x})$ then there exists $\lambda>0$ and $\boldsymbol{z}\in\mathcal{T}(\boldsymbol{x})\cap S^{n-1}$ such that $\boldsymbol{y}=\lambda\boldsymbol{z}$. So let $\boldsymbol{z}\in\mathcal{T}(\boldsymbol{x})\cap S^{n-1}$. Then, for all $\lambda>0$ we have that $\|A(\lambda\boldsymbol{z})\|_2>0$. This means that $(\mathcal{T}(\boldsymbol{x})\setminus\{\boldsymbol{0}\})\cap Ker(A)=\emptyset$ and thus $Ker(A)\cap\mathcal{T}(\boldsymbol{x})=\{\boldsymbol{0}\}$. □

The next theorem contains a sufficient condition for robust recovery in the noisy case.

Theorem 5.3.2 (Sufficient condition for successful recovery-noisy case)**.** *Let $\boldsymbol{x}\in\mathbb{R}^n$ and $A\in\mathbb{R}^{m\times n}$. Then for a solution $\hat{\boldsymbol{x}}$ of optimization task 2.9.2, with $\boldsymbol{y}=A\boldsymbol{x}+\boldsymbol{e}$, $\|\boldsymbol{e}\|_2\leq\epsilon$, it holds that*

$$\|\hat{\boldsymbol{x}}-\boldsymbol{x}\|_2\leq\frac{2\epsilon}{\tau},$$

for some $\tau>0$, if

$$\inf_{\boldsymbol{z}\in\mathcal{T}(\boldsymbol{x})\cap S^{n-1}}\|A\boldsymbol{z}\|_2\geq\tau.$$

Proof. Let $\hat{\boldsymbol{x}}$ be a solution of (2.9.2) with $\boldsymbol{y}=A\boldsymbol{x}+\boldsymbol{e}$, then $\|\hat{\boldsymbol{x}}\|_1\leq\|\boldsymbol{x}\|_1$. From the definition of the tangent cone follows that $(\hat{\boldsymbol{x}}-\boldsymbol{x})\in\mathcal{T}(\boldsymbol{x})$ and thus

$$\boldsymbol{z}=\frac{\hat{\boldsymbol{x}}-\boldsymbol{x}}{\|\hat{\boldsymbol{x}}-\boldsymbol{x}\|_2}\in\mathcal{T}(\boldsymbol{x}),$$

assuming that $\hat{\boldsymbol{x}}\neq\boldsymbol{x}$. It holds that $\|\boldsymbol{z}\|_2=1$, hence

$$\|A\boldsymbol{z}\|_2>\tau\Rightarrow\|A(\hat{\boldsymbol{x}}-\boldsymbol{x})\|_2\geq\tau\|\hat{\boldsymbol{x}}-\boldsymbol{x}\|_2.$$

So we can write that

$$\begin{aligned}\tau\|\hat{\boldsymbol{x}}-\boldsymbol{x}\|_2 \leq \|A(\hat{\boldsymbol{x}}-\boldsymbol{x}\|_2 \leq \\ \leq \|A\hat{\boldsymbol{x}}-\boldsymbol{y}\|_2 + \|A\boldsymbol{x}-\boldsymbol{y})\|_2 \leq \\ \leq \boldsymbol{e}+\boldsymbol{e} \leq 2\epsilon \Rightarrow \\ \Rightarrow \|\hat{\boldsymbol{x}}-\boldsymbol{x}\|_2 \leq \frac{2\epsilon}{\tau}.\end{aligned}$$

□

In the previous definitions we referred to some arbitrary matrix $A \in \mathbb{R}^{m\times n}$. The focus in this chapter is on random matrices, hence it is apparent by now that we want an estimate of the following quantity

$$\mathbb{P}\left(\inf_{\boldsymbol{z}\in T(\boldsymbol{x})\cap S^{n-1}} \|A\boldsymbol{z}\|_2 < t\right),\ t>0.$$

Roughly, we want to find the probability that the kernel of the measurement matrix A misses the tangent cone at a point $\boldsymbol{x}$. If we put it in a more general framework, we want to obtain an estimate of the probability that a random subspace (following a uniform distribution) misses a set. The result we are looking for is called *Gordon's escape through the mesh theorem.* [Gor88]

Theorem 5.3.3. *(Gordon's escape through the mesh theorem) Let $A \in \mathbb{R}^{m\times n}$ be a Gaussian random matrix and $T \subseteq S^{n-1}$ be a set. Then,*

$$\mathbb{P}\left(\inf_{\boldsymbol{z}\in T} \|A\boldsymbol{z}\|_2 \leq \boldsymbol{g}_m - w(T) - t\right) \leq e^{-t^2/2},\ t>0,$$

where $\boldsymbol{g}_m = \mathbb{E}\left[\|\boldsymbol{g}\|_2\right]$ and $w(T)$ is the Gaussian width of the set T.

Proof. We define the following Gaussian processes

$$X_{\boldsymbol{x},\boldsymbol{y}} = \langle A\boldsymbol{x},\boldsymbol{y}\rangle = \sum_{i=1}^{m}\sum_{j=1}^{n} A_{ij}x_j y_i$$

and

$$Y_{\boldsymbol{x},\boldsymbol{y}} = \langle \boldsymbol{g},\boldsymbol{x}\rangle + \langle \boldsymbol{h},\boldsymbol{y}\rangle,$$

where $\boldsymbol{g} \in \mathbb{R}^n$ and $\boldsymbol{h} \in \mathbb{R}^m$ are independent standard Gaussian random vectors.

Let $\boldsymbol{x},\boldsymbol{x}' \in S^{n-1}$ and $\boldsymbol{y},\boldsymbol{y}' \in S^{m-1}$. Then we have that

$$\begin{aligned}\mathbb{E}\left[|X_{\boldsymbol{x},\boldsymbol{y}} - X_{\boldsymbol{x}',\boldsymbol{y}'}|^2\right] &= \mathbb{E}\left[\left|\sum_{i=1}^{m}\sum_{j=1}^{n} A_{ij}(x_j y_i - x'_j y'_i)\right|^2\right] = \sum_{i=1}^{m}\sum_{j=1}^{n}(x_j y_i - x'_j y'_i)^2 = \\ &= \sum_{i=1}^{m}\sum_{j=1}^{n}\left(x_j^2 y_i^2 - 2x_j y_i x'_j y'_i + (x'_j)^2(y'_i)^2\right) = \\ &= \left(\sum_{j=1}^{n} x_j^2\right)\left(\sum_{i=1}^{m} y_i^2\right) - 2\left(\sum_{i=1}^{m} y_i y'_i\right)\left(\sum_{j=1}^{n} x_j x'_j\right) + \left(\sum_{j=1}^{n}(x'_j)^2\right)\left(\sum_{i=1}^{m}(y'_i)^2\right) = \\ &= \|\boldsymbol{x}\|_2^2\|\boldsymbol{y}\|_2^2 - 2\langle \boldsymbol{x},\boldsymbol{x}'\rangle\langle \boldsymbol{y},\boldsymbol{y}'\rangle + \|\boldsymbol{x}'\|_2^2\|\boldsymbol{y}'\|_2^2 \\ &= 2 - 2\langle \boldsymbol{x},\boldsymbol{x}'\rangle\langle \boldsymbol{y},\boldsymbol{y}'\rangle.\end{aligned}$$

We have used the fact that A is a Gaussian random matrix and as a result for its entries we can write

$$\mathbb{E}\left[A_{ij}A_{kl}\right] = \begin{cases} 0 & ,if\, ij \neq kl \\ 1 & ,if\, ij = kl \end{cases}.$$

Also, for the process $Y_{\boldsymbol{x},\boldsymbol{y}}$ we have that

$$\begin{aligned}
\mathbb{E}\left[\left|Y_{\boldsymbol{x},\boldsymbol{y}} - Y_{\boldsymbol{x}',\boldsymbol{y}'}\right|^2\right] &= \mathbb{E}\left[\left|\langle \boldsymbol{g},\boldsymbol{x}\rangle + \langle \boldsymbol{h},\boldsymbol{y}\rangle - \langle \boldsymbol{g},\boldsymbol{x}'\rangle - \langle \boldsymbol{h},\boldsymbol{y}'\rangle\right|^2\right] = \\
&= \mathbb{E}\left[\left|\langle \boldsymbol{g},\boldsymbol{x}-\boldsymbol{x}'\rangle + \langle \boldsymbol{h},\boldsymbol{y}-\boldsymbol{y}'\rangle\right|^2\right] = \\
&= \mathbb{E}\left[\left|\langle \boldsymbol{g},\boldsymbol{x}-\boldsymbol{x}'\rangle\right|^2\right] + \mathbb{E}\left[\left|\langle \boldsymbol{h},\boldsymbol{y}-\boldsymbol{y}'\rangle\right|^2\right] = \qquad \text{(Independence of } g,h\text{)} \\
&= \|\boldsymbol{x}-\boldsymbol{x}'\|_2^2 + \|\boldsymbol{y}-\boldsymbol{y}'\|_2^2 \qquad \text{(Isotropicity),(lemma 5.2.1)} \\
&= \|\boldsymbol{x}\|_2^2 - 2\langle \boldsymbol{x},\boldsymbol{x}'\rangle + \|\boldsymbol{x}'\|_2^2 + \|\boldsymbol{y}\|_2^2 - 2\langle \boldsymbol{y},\boldsymbol{y}'\rangle + \|\boldsymbol{y}'\|_2^2 = \\
&= 4 - 2\langle \boldsymbol{x},\boldsymbol{x}'\rangle - 2\langle \boldsymbol{y},\boldsymbol{y}'\rangle .
\end{aligned}$$

Working the difference between the previous two expressions yields.

$$\begin{aligned}
\mathbb{E}\left[\left|Y_{\boldsymbol{x},\boldsymbol{y}} - Y_{\boldsymbol{x}',\boldsymbol{y}'}\right|^2\right] - \mathbb{E}\left[\left|X_{\boldsymbol{x},\boldsymbol{y}} - X_{\boldsymbol{x}',\boldsymbol{y}'}\right|^2\right] &= 2\left(1 - \langle \boldsymbol{x},\boldsymbol{x}'\rangle - \langle \boldsymbol{y},\boldsymbol{y}'\rangle + \langle \boldsymbol{x},\boldsymbol{x}'\rangle\langle \boldsymbol{y},\boldsymbol{y}'\rangle\right) = \\
&= 2(1 - \langle \boldsymbol{x},\boldsymbol{x}'\rangle)(1 - \langle \boldsymbol{y},\boldsymbol{y}'\rangle).
\end{aligned}$$

Cauchy-Schwartz inequality and the fact that $\boldsymbol{x},\boldsymbol{x}' \in S^{n-1}$, $\boldsymbol{y},\boldsymbol{y}' \in S^{m-1}$ yield

$$\langle \boldsymbol{x},\boldsymbol{x}'\rangle \leq \|\boldsymbol{x}\|_2\|\boldsymbol{x}'\|_2 = 1$$

and

$$\langle \boldsymbol{y},\boldsymbol{y}'\rangle \leq \|\boldsymbol{y}\|_2\|\boldsymbol{y}'\|_2 = 1.$$

As a result,

$$\mathbb{E}\left[\left|Y_{\boldsymbol{x},\boldsymbol{y}} - Y_{\boldsymbol{x}',\boldsymbol{y}'}\right|^2\right] - \mathbb{E}\left[\left|X_{\boldsymbol{x},\boldsymbol{y}} - X_{\boldsymbol{x}',\boldsymbol{y}'}\right|^2\right] \geq 0.$$

Note that

$$\mathbb{E}\left[\left|Y_{\boldsymbol{x},\boldsymbol{y}} - Y_{\boldsymbol{x}',\boldsymbol{y}'}\right|^2\right] - \mathbb{E}\left[\left|X_{\boldsymbol{x},\boldsymbol{y}} - X_{\boldsymbol{x}',\boldsymbol{y}'}\right|^2\right] = 0 \Leftrightarrow \langle \boldsymbol{x},\boldsymbol{x}'\rangle = 1 \text{ or } \langle \boldsymbol{y},\boldsymbol{y}'\rangle = 1.$$

So, he have established that

$$\begin{cases} \mathbb{E}\left[\left|X_{\boldsymbol{x},\boldsymbol{y}} - X_{\boldsymbol{x}',\boldsymbol{y}'}\right|^2\right] \leq \mathbb{E}\left[\left|Y_{\boldsymbol{x},\boldsymbol{y}} - Y_{\boldsymbol{x}',\boldsymbol{y}'}\right|^2\right] & ,if\, \boldsymbol{x} \neq \boldsymbol{x}' \\ \mathbb{E}\left[\left|X_{\boldsymbol{x},\boldsymbol{y}} - X_{\boldsymbol{x}',\boldsymbol{y}'}\right|^2\right] = \mathbb{E}\left[\left|Y_{\boldsymbol{x},\boldsymbol{y}} - Y_{\boldsymbol{x}',\boldsymbol{y}'}\right|^2\right] & ,if\, \boldsymbol{x} = \boldsymbol{x}' \end{cases}$$

Using Gordon's lemma (theorem 4.7.1) and the observation that states its generalization in Gaussian processes we get

$$\begin{aligned}
\mathbb{E}\left[\inf_{\boldsymbol{x}\in T}\max_{\boldsymbol{y}\in S^{m-1}} X_{\boldsymbol{x},\boldsymbol{y}}\right] &\geq \mathbb{E}\left[\inf_{\boldsymbol{x}\in T}\max_{\boldsymbol{y}\in S^{m-1}} Y_{\boldsymbol{x},\boldsymbol{y}}\right] = \\
&= \mathbb{E}\left[\inf_{\boldsymbol{x}\in T}\max_{\boldsymbol{y}\in S^{m-1}} \{\langle \boldsymbol{g},\boldsymbol{x}\rangle + \langle \boldsymbol{h},\boldsymbol{y}\rangle\}\right] = \\
&= \mathbb{E}\left[\inf_{\boldsymbol{x}\in T}\left\{\langle \boldsymbol{g},\boldsymbol{x}\rangle + \max_{\boldsymbol{y}\in S^{m-1}} \langle \boldsymbol{h},\boldsymbol{y}\rangle\right\}\right] = \\
&= \mathbb{E}\left[\inf_{\boldsymbol{x}\in T}\{\langle \boldsymbol{g},\boldsymbol{x}\rangle + \|\boldsymbol{h}\|_2\}\right] = \\
&= \mathbb{E}\left[\|\boldsymbol{h}\|_2\right] + \mathbb{E}\left[\inf_{\boldsymbol{x}\in T}\langle \boldsymbol{g},\boldsymbol{x}\rangle\right] = \qquad \text{(symmetry of std. Gaussian vector)} \\
&= \mathbb{E}\left[\|\boldsymbol{h}\|_2\right] - \mathbb{E}\left[\sup_{\boldsymbol{x}\in T}\langle \boldsymbol{g},\boldsymbol{x}\rangle\right] = \\
&= \boldsymbol{h}_m - w(T).
\end{aligned}$$

Notice that

$$\mathbb{E}\left[\inf_{\boldsymbol{x}\in T}\|A\boldsymbol{x}\|_2\right] = \mathbb{E}\left[\inf_{\boldsymbol{x}\in T}\max_{\boldsymbol{y}\in S^{m-1}} X_{\boldsymbol{x},\boldsymbol{y}}\right].$$

Therefore, combining the previous two expressions yields

$$\mathbb{E}\left[\inf_{\boldsymbol{x}\in T}\|A\boldsymbol{x}\|_2\right] \geq \boldsymbol{h}_m - w(T). \tag{5.11}$$

This result is known as *Gordon's Comparison theorem.*

We define the function

$$F(A) = \inf_{\boldsymbol{x}\in T}\|A\boldsymbol{x}\|_2.$$

The function F is a Lipschitz function with respect to the Frobenius norm and constant 1. Indeed,

$$\begin{aligned}
\inf_{\boldsymbol{x}\in T}\{\|A\boldsymbol{x}\|_2\} &= \inf_{\boldsymbol{x}\in T}\{\|(A-B)\boldsymbol{x}+B\boldsymbol{x}\|_2\} \leq \inf_{\boldsymbol{x}\in T}\{\|(A-B)\boldsymbol{x}\|_2 + \|B\boldsymbol{x}\|_2\} \leq \\
&\leq \inf_{\boldsymbol{x}\in T}\{\|A-B\|_{2\to 2}\|\boldsymbol{x}\|_2\} + \inf_{\boldsymbol{x}\in T}\{\|B\boldsymbol{x}\|_2\} \leq \qquad (T\subseteq S^{n-1}) \\
&\leq \|(A-B)\|_{2\to 2} + \inf_{\boldsymbol{x}\in T}\{\|B\boldsymbol{x}\|_2\} \leq \\
&\leq \|(A-B)\|_F + \inf_{\boldsymbol{x}\in T}\{\|B\boldsymbol{x}\|_2\} \qquad (\textit{Lemma } A.1.1) \\
&\Rightarrow \inf_{\boldsymbol{x}\in T}\{\|A\boldsymbol{x}\|_2\} - \inf_{\boldsymbol{x}\in T}\{\|B\boldsymbol{x}\|_2\} \leq \|(A-B)\|_F.
\end{aligned}$$

In the same way we can show that

$$\inf_{\boldsymbol{x}\in T}\{\|B\boldsymbol{x}\|_2\} - \inf_{\boldsymbol{x}\in T}\{\|A\boldsymbol{x}\|_2\} \leq \|(A-B)\|_F$$

So we conclude that,

$$|F(A) - F(B)| \leq \|(A-B)\|_F.$$

Notice that F is Lipschitz with respect to the Frobenius norm (which is the l_2 norm if we treat the matrix as a vector) with constant $L = 1$ and A is a standard Gaussian vector in $\mathbb{R}^{n\cdot m}$ if we treat matrix A as a vector. Consequently, the function F satisfies all the conditions of theorem 4.8.3 and thus

$$\mathbb{P}\left[\inf_{\boldsymbol{x}\in T}\|A\boldsymbol{x}\|_2 \leq \mathbb{E}\left[\inf_{\boldsymbol{x}\in T}\|A\boldsymbol{x}\|_2\right] - t\right] \leq e^{-t^2/2}.$$

The use of Gordon's comparison theorem (5.11) leads to the final result,

$$\mathbb{P}\left[\inf_{\boldsymbol{x}\in T}\|A\boldsymbol{x}\|_2 \leq \boldsymbol{h}_m - w(T) - t\right] \leq \mathbb{P}\left[\inf_{\boldsymbol{x}\in T}\|A\boldsymbol{x}\|_2 \leq \mathbb{E}\left[\inf_{\boldsymbol{x}\in T}\|A\boldsymbol{x}\|_2\right] - t\right] \leq e^{-t^2/2}.$$

□

At this point we have at our disposal a bound for the probability that a random (uniformly distributed) subspace misses a subset of the $n-1$ dimensional sphere. In order to adapt this result to the specific situation we are facing we need to specify the two parameters it employs, the expectation of the l_2 norm of a standard Gaussian random vector $\boldsymbol{g}_m$ and the Gaussian width of the set $T \cap S^{n-1}$. We have already calculated an estimate for $\boldsymbol{g}_m$ (section 4.5, theorem 4.5.1), which is sufficient. Hence, we need an estimate for the Gaussian width of the set of unit-norm vectors of the tangent cone at a sparse vector $x \in \Sigma_k$. Before we proceed, we need to define the *normal cone of the l_1 norm*.

Definition 5.3.2 (Normal cone of l_1 norm)**.**

$$\mathcal{N}(\boldsymbol{x}) = \{\boldsymbol{z} \in \mathbb{R}^n : \forall \boldsymbol{w}\, s.t\, \|\boldsymbol{w}\|_1 \leq \|\boldsymbol{x}\|_1 \;\; \textit{it holds that} \;\; \langle \boldsymbol{z}, \boldsymbol{w} - \boldsymbol{x}\rangle \leq 0\}\,. \quad (5.12)$$

So we can now state the following proposition.

Proposition 5.3.2. *Let $\boldsymbol{g} \in \mathbb{R}^n$ is be a standard Gaussian random vector. Then we have that*

$$w\left(\mathcal{T}(\boldsymbol{x}) \cap S^{n-1}\right) \leq \mathbb{E}\left[\min_{\boldsymbol{z}\in\mathcal{N}(\boldsymbol{x})}\|\boldsymbol{g} - \boldsymbol{z}\|_2\right].$$

Proof. We have that

$$\begin{aligned} w\left(\mathcal{T}(\boldsymbol{x}) \cap S^{n-1}\right) = \mathbb{E}\left[\max_{\mathcal{T}(\boldsymbol{x})\cap S^{n-1}} \langle \boldsymbol{g}, \boldsymbol{z}\rangle\right] &\leq \mathbb{E}\left[\max_{\boldsymbol{z}\in\mathcal{T}(\boldsymbol{x}),\|\boldsymbol{z}\|_2\leq 1} \langle \boldsymbol{g}, \boldsymbol{z}\rangle\right] \leq \quad \text{(Lemma A.2.1)} \\ &\leq \mathbb{E}\left[\min_{\boldsymbol{z}\in\mathcal{T}^\circ(\boldsymbol{x})}\|\boldsymbol{g} - \boldsymbol{z}\|_2\right] = \mathbb{E}\left[\min_{\boldsymbol{z}\in\mathcal{N}(\boldsymbol{x})}\|\boldsymbol{g} - \boldsymbol{z}\|_2\right]. \end{aligned}$$

In the last equality we have used the fact that the polar cone of the tangent cone is the normal cone, i.e $\mathcal{T}^\circ(\boldsymbol{x}) = \mathcal{N}(\boldsymbol{x})$. (def. A.2.2) □

It is apparent that in order to obtain a bound for the Gaussian width we need to calculate the normal cone of the l_1 norm at a sparse vector. The following theorem provides that result.

Lemma 5.3.1. *Let $\boldsymbol{x} \in \mathbb{R}^n$ such that $S = supp(\boldsymbol{x}) \subseteq \{1, 2, \dots, n\}$. Then the normal cone on $\boldsymbol{x}$ can be written as*

$$\mathcal{N}(\boldsymbol{x}) = \bigcup_{t \geq 0} \left\{ \boldsymbol{z} \in \mathbb{R}^n : \begin{cases} z_i = t sgn(x_i) & , for\, i \in S \\ |z_i| < t & , for\, i \in \overline{S} \end{cases} \right\} \tag{5.13}$$

Proof. First, we are going to show that the right-hand side of (5.13) is a subset of the right-hand side of (5.12). So, let $\boldsymbol{z} \in \mathbb{R}^n$ be an element of the right-hand side of (5.13). Also we consider an arbitrary $\boldsymbol{y}$ such that $\|\boldsymbol{y}\|_1 \leq \|\boldsymbol{x}\|_1$. Then

$$\begin{aligned} \langle \boldsymbol{z}, \boldsymbol{y} - \boldsymbol{x} \rangle = \langle \boldsymbol{z}, \boldsymbol{y} \rangle - \langle \boldsymbol{z}, \boldsymbol{x} \rangle \leq \|\boldsymbol{z}\|_\infty \|\boldsymbol{y}\|_1 - \|\boldsymbol{z}\|_\infty \|\boldsymbol{x}\|_1 = \\ = \|\boldsymbol{z}\|_\infty (\|\boldsymbol{y}\|_1 - \|\boldsymbol{x}\|_1) \leq 0. \end{aligned}$$

Therefore, $\boldsymbol{z} \in \mathcal{N}(\boldsymbol{x})$.

Now we are going to show the converse direction, i.e the right-hand side of (5.12) is a subset of the right-hand side of (5.13). So, let $\boldsymbol{z} \in \mathcal{N}(\boldsymbol{x})$, i.e for all $\boldsymbol{y}$ such that $\|\boldsymbol{y}\|_1 \leq \|\boldsymbol{x}\|_1$ we have that $\langle \boldsymbol{z}, \boldsymbol{y} - \boldsymbol{x} \rangle \leq 0$. Then we choose a vector $\boldsymbol{y}$ such that $\|\boldsymbol{y}\|_1 = \|\boldsymbol{x}\|_1$ and

$$\boldsymbol{y} = \begin{cases} y_i & , for\, i \,\text{s.t}\, |z_i| = \|\boldsymbol{z}\|_\infty \\ 0 & , otherwise \end{cases},$$

with $sgn(y_i) = sgn(z_i)$. So we have that

$$\|\boldsymbol{z}\|_\infty \|\boldsymbol{y}\|_1 = \langle \boldsymbol{z}, \boldsymbol{y} \rangle \leq \langle \boldsymbol{z}, \boldsymbol{x} \rangle \leq \|\boldsymbol{z}\|_\infty \|\boldsymbol{x}\|_1 = \|\boldsymbol{z}\|_\infty \|\boldsymbol{y}\|_1.$$

The previous expression implies that

$$\langle \boldsymbol{z}, \boldsymbol{x} \rangle = \|\boldsymbol{z}\|_\infty \|\boldsymbol{x}\|_1,$$

which in turn implies that $z_i = sgn(x_i)\|\boldsymbol{z}\|_\infty, \forall i \in S$. Also, notice that $|z_i| \leq \|\boldsymbol{z}\|_\infty$, $\forall i \in \overline{S}$. Therefore, if we choose $t = \|\boldsymbol{z}\|_\infty$ we can see that $\boldsymbol{z}$ belongs to the right-hand side of (5.13). □

Now we are ready to provide a bound for the Gaussian width of the set of unit-norm vectors of the tangent cone at a sparse vector.

Proposition 5.3.3. *Let $\boldsymbol{x} \in \mathbb{R}^n$ be a k-sparse vector. Then we have that*

$$\left(w\left(\mathcal{T}(\boldsymbol{x}) \cap S^{n-1}\right)\right)^2 \leq 2k\, ln\left(\frac{en}{k}\right).$$

Proof. First, notice that we can attain the following result

$$\begin{aligned} \left(w\left(\mathcal{T}(\boldsymbol{x}) \cap S^{n-1}\right)\right)^2 \leq \left(\mathbb{E}\left[\min_{\boldsymbol{z} \in \mathcal{N}(\boldsymbol{x})} \|\boldsymbol{g} - \boldsymbol{z}\|_2\right]\right)^2 \leq \qquad \text{(proposition 5.3.2)} \\ \leq \mathbb{E}\left[\min_{\boldsymbol{z} \in \mathcal{N}(\boldsymbol{x})} \|\boldsymbol{g} - \boldsymbol{z}\|_2^2\right]. \qquad \text{(Holder's inequality)} \end{aligned}$$

Let $S = supp(\boldsymbol{x})$ be the support of $\boldsymbol{x}$. We isolate the expression inside the expectation which yields,

$$\min_{\boldsymbol{z}\in\mathcal{N}(\boldsymbol{x})}\|\boldsymbol{g}-\boldsymbol{z}\|_2^2 = \min_{\substack{t\geq 0\\ |z_i|\leq t, i\in\overline{S}}}\left\{\sum_{i\in S}\left(g_i - tsgn(x_i)\right)^2 + \sum_{i\in\overline{S}}\left(g_i - z_i\right)^2\right\} = \qquad \text{(lemma 5.3.1)}$$

$$= \min_{t\geq 0}\left\{\sum_{i\in S}\left(g_i - tsgn(x_i)\right)^2 + \min_{|z_i|\leq t, i\in\overline{S}}\left\{\sum_{i\in\overline{S}}\left(g_i - z_i\right)^2\right\}\right\}.$$

It is easy to verify that

$$\min_{|z_i|\leq t}\left(g_i - z_i\right)^2 = S_t^2(g_i),$$

where S_t is the soft thresholding operator

$$S_t(u) = \begin{cases} x+t & , x\leq -t \\ 0 & , -t\leq u\leq t \\ x-t & , x\geq t \end{cases}.$$

Then,

$$\min_{\boldsymbol{z}\in\mathcal{N}(\boldsymbol{x})}\|\boldsymbol{g}-\boldsymbol{z}\|_2^2 = \min_{t\geq 0}\left\{\sum_{i\in S}\left(g_i - tsgn(x_i)\right)^2 + \sum_{i\in\overline{S}} S_t^2(g_i)\right\}.$$

Thus, for some fixed $t>0$, independent from $\boldsymbol{g}$ we have that

$$\begin{aligned}
\mathbb{E}\left[\min_{\boldsymbol{z}\in\mathcal{N}(\boldsymbol{x})}\|\boldsymbol{g}-\boldsymbol{z}\|_2^2\right] &\leq \mathbb{E}\left[\sum_{i\in S}\left(g_i - tsgn(x_i)\right)^2 + \sum_{i\in\overline{S}} S_t^2(g_i)\right] = \\
&= \mathbb{E}\left[\sum_{i\in S}\left(g_i - tsgn(x_i)\right)^2\right] + \mathbb{E}\left[\sum_{i\in\overline{S}} S_t^2(g_i)\right] \leq \\
&\leq k\mathbb{E}\left[(g+t)^2\right] + \sum_{i\in\overline{S}}\mathbb{E}\left[S_t^2(g_i)\right] = \qquad (g \text{ standard Gaussian r.var}) \\
&= k(1+t^2) + (n-k)\mathbb{E}\left[S_t^2(g_i)\right].
\end{aligned}$$

For the expectation of the square of the soft thresholding operator we have that

$$\mathbb{E}\left[S_t^2\right] = \frac{1}{\sqrt{2\pi}}\left(\int_{-\infty}^{-t} S_t^2(u)e^{-u^2/2}du + \int_{t}^{+\infty} S_t^2(u)e^{-u^2/2}du\right) = \qquad \text{(Symmetry of } \boldsymbol{g} \text{ and } S_t\text{)}$$

$$= \frac{2}{\sqrt{2\pi}}\int_{t}^{+\infty}(u-t)^2 e^{-u^2/2}du = \frac{2}{\sqrt{2\pi}}\int_{0}^{+\infty} r^2 e^{-(r+t)^2/2}dr = \qquad (r = u - t)$$

$$= \frac{2}{\sqrt{2\pi}}e^{-t^2/2}\int_{0}^{+\infty} r^2 e^{-r^2/2}e^{-rt}dr \leq \frac{2}{\sqrt{2\pi}}e^{-t^2/2}\int_{0}^{+\infty} r^2 e^{-r^2/2}dr =$$

$$= e^{-t^2/2}\mathbb{E}\left[g^2\right] = e^{-t^2/2} \qquad (\boldsymbol{g} \text{ standard Gaussian r.var})$$

Hence, using the previous two expressions yields

$$\left(w\left(\mathcal{T}(\boldsymbol{x}) \cap S^{n-1}\right)\right)^2 \leq \min_{t\geq 0}\left\{k(1+t^2) + (n-k)e^{-t^2/2}\right\} \leq$$
$$\leq \min_{t\geq 0}\left\{k(1+t^2) + ne^{-t^2/2}\right\}.$$

Finally, if we pick $t = \sqrt{2ln\left(\frac{n}{k}\right)}$ we can obtain

$$\left(w\left(\mathcal{T}(\boldsymbol{x}) \cap S^{n-1}\right)\right)^2 \leq \min_{t\geq 0}\left\{k(1+t^2) + ne^{-t^2/2}\right\} \leq$$
$$\leq k(1+2ln\left(\frac{n}{k}\right)) + k =$$
$$= 2kln\left(\frac{en}{k}\right).$$

□

Now we have all the tools that are necessary in order to state the two main theorems concerning non-uniform recovery with Gaussian matrices.

Theorem 5.3.4 (Non-uniform recovery with Gaussian matrices-noiseless case)**.** *Let $\boldsymbol{x} \in \mathbb{R}^n$ be a k-sparse vector, $A \in \mathbb{R}^{m\times n}$ be a random Gaussian matrix and*

$$\frac{m^2}{m+1} \geq 2k\left(\sqrt{ln\left(\frac{en}{k}\right)} + \sqrt{\frac{ln(\epsilon^{-1})}{k}}\right)^2, \tag{5.14}$$

for some $\epsilon \in (0,1)$. Then, with probability at least $1-\epsilon$, $\boldsymbol{x}$ is the unique solution of the optimization task (2.6.3), with $\boldsymbol{y} = A\boldsymbol{x}$.

Proof. It is easy to see that

$$\mathbb{P}\left(\min_{\boldsymbol{z}\in\mathcal{T}(\boldsymbol{x})\cap S^{n-1}}\|A\boldsymbol{z}\|_2 > 0\right) \geq$$
$$\geq \mathbb{P}\left(\min_{\boldsymbol{z}\in\mathcal{T}(\boldsymbol{x})\cap S^{n-1}}\|A\boldsymbol{z}\|_2 > \boldsymbol{g}_m - w(\mathcal{T}(\boldsymbol{x})\cap S^{n-1}) - t\right) =$$
$$= 1 - \mathbb{P}\left(\min_{\boldsymbol{z}\in\mathcal{T}(\boldsymbol{x})\cap S^{n-1}}\|A\boldsymbol{z}\|_2 \leq \boldsymbol{g}_m - w(\mathcal{T}(\boldsymbol{x})\cap S^{n-1}) - t\right)$$

provided that

$$\boldsymbol{g}_m - w(\mathcal{T}(\boldsymbol{x}) \cap S^{n-1}) - t \geq 0.$$

Notice that from theorem 4.5.1 we know that

$$\boldsymbol{g}_m \geq \frac{m}{\sqrt{m+1}}$$

and from proposition 5.3.3 we can obtain the following bound

$$w\left(\mathcal{T}(\boldsymbol{x}) \cap S^{n-1}\right) \leq \sqrt{2k\, ln\left(\frac{en}{k}\right)}.$$

Hence, if we set $t = \sqrt{2ln(2\epsilon^{-1})}$ we have that

$$\begin{aligned} &\boldsymbol{g}_m - w(\mathcal{T}(\boldsymbol{x}) \cap S^{n-1}) - t \geq 0 \Rightarrow \\ &\frac{m^2}{m+1} \geq 2k \left(\sqrt{ln\left(\frac{en}{k}\right)} + \sqrt{\frac{ln(\epsilon^{-1})}{k}} \right)^2 \end{aligned}$$

and

$$\begin{aligned} &\mathbb{P}\left(\min_{\boldsymbol{z} \in \mathcal{T}(\boldsymbol{x}) \cap S^{n-1}} \|A\boldsymbol{z}\|_2 > 0 \right) \geq \\ &\geq 1 - e^{-t^2/2} = 1 - \epsilon. \end{aligned}$$

We conclude that if expression (5.14) holds, then proposition 5.3.1 establishes that we have successful recovery with probability at least $1 - \epsilon$.

□

In the noisy case we can obtain the following result.

Theorem 5.3.5 (Non-uniform recovery with Gaussian matrices-noisy case)**.** *Let $\boldsymbol{x} \in \mathbb{R}^n$ be a k-sparse vector, $A \in \mathbb{R}^{m \times n}$ be a random Gaussian matrix and*

$$\frac{m^2}{m+1} \geq 2k \left(\sqrt{ln\left(\frac{en}{k}\right)} + \sqrt{\frac{ln(\epsilon^{-1})}{k}} + \frac{\tau}{\sqrt{k}} \right)^2$$

for some $\epsilon \in (0,1)$ and $\tau > 0$. Then for every solution $\hat{\boldsymbol{x}}$ of the optimization task (2.9.1), with $\boldsymbol{y} = A\boldsymbol{x} + \boldsymbol{e}$, $\|\boldsymbol{e}\|_2 < \eta$, it holds that

$$\|\boldsymbol{x} - \hat{\boldsymbol{x}}\|_2 \leq \frac{2\eta}{\tau},$$

with probability at least $1 - \epsilon$.

Proof. Using the same reasoning with the previous proof, notice that if we set $t = \sqrt{2ln(2\epsilon^{-1})}$ we can deduce that

$$\begin{aligned} &\boldsymbol{g}_m - w(\mathcal{T}(\boldsymbol{x}) \cap S^{n-1}) - \tau - t \geq 0 \Rightarrow \\ &\frac{m^2}{m+1} \geq 2k \left(\sqrt{ln\left(\frac{en}{k}\right)} + \sqrt{\frac{ln(\epsilon^{-1}}{k}} + \frac{\tau}{\sqrt{k}} \right)^2. \end{aligned}$$

Using theorem (5.3.2) and the expression below we complete the proof

$$\begin{aligned}
&\mathbb{P}\left(\min_{\boldsymbol{z}\in\mathcal{T}(\boldsymbol{x})\cap S^{n-1}}\|A\boldsymbol{z}\|_2 > \tau\right) \geq \\
&\geq \mathbb{P}\left(\min_{\boldsymbol{z}\in\mathcal{T}(\boldsymbol{x})\cap S^{n-1}}\|A\boldsymbol{z}\|_2 > \boldsymbol{g}_m - w(\mathcal{T}(\boldsymbol{x})\cap S^{n-1}) - \tau - t\right) \geq \\
&\geq 1 - e^{-t^2/2} = 1 - \epsilon.
\end{aligned}$$

□

Part III

Algorithms

Chapter 6

Algorithms for matrix completion

The purpose of this chapter is to introduce five different approaches for the matrix completion problem and evaluate the performance of the respective algorithms on synthetic and real data (MovieLens dataset). The experiments were performed in MATLAB.

6.1 Algorithms

In this section we will describe the five algorithms that we are going to evaluate at the next section. Before we do that we will make some general remarks and some comments about the notation we are going to use.

We denote the matrices with capital letters, such as X, the ith column of a matrix X as $X_{(\cdot,i)}$ and the ith row as $X_{(i,\cdot)}$. We consider $X_{(\cdot,i)}$ to be a column vector and $X_{(i,\cdot)}$ to be a row vector. Note that we are also going to use the notation $\boldsymbol{x}_i$ for the ith row of X and $\boldsymbol{x}_{c,i}$ for the ith column of X, but these will be clearly stated. Finally, the expression $C = diag(X_{(i,\cdot)})$ denotes a diagonal matrix, whose entries are the elements of the row vector $X_{(i,\cdot)}$.

We introduce the operator P_Ω for a sampling set $\Omega \subseteq \{1,\dots,k\} \times \{1,\dots,n\}$ (the positions of the known entries) and a matrix $X \in \mathbb{R}^{k\times n}$, which is defined as follows.

$$P_\Omega(X) = \begin{cases} x_{ij} & ,(i,j) \in \Omega \\ 0 & ,(i,j) \notin \Omega \end{cases}$$

Unless otherwise stated, we consider the cardinality of the sampling set, i.e $card(\Omega)$, to be m.

Consider an incomplete matrix $X \in \mathbb{R}^{k\times n}$. We denote with $P \in \mathbb{R}^{k\times n}$ the matrix with one in the positions (i,j) of the known entries and zero otherwise, i.e.

$$P = \begin{cases} 1 & ,(i,j) \in \Omega \\ 0 & ,(i,j) \notin \Omega \end{cases},$$

where Ω is the sampling set of X.

6.1.1 Proximal forward-backward splitting

In chapter 3 we introduced the nuclear norm minimization task (definition (3.4.3)). We can relax the constraints of this task and express the cost function in a different form. The new cost function is

$$\boxed{f_1(X) = \frac{1}{2}\|P_\Omega(M) - P_\Omega(X)\|_F^2 + \lambda\|X\|_*}, \tag{6.1}$$

for some $\lambda > 0$.

The respective minimization problem is provided below.

Definition 6.1.1 (Relaxed nuclear norm minimization)**.** *Let $M \in \mathbb{R}^{k\times n}$. The nuclear norm minimization task is*

$$\min_{X\in\mathbb{R}^{k\times n}} \frac{1}{2}\|P_\Omega(M) - P_\Omega(X)\|_F^2 + \lambda\|X\|_*.$$

We can interpret the previous minimization problem as the task that tries to minimize the first term of (6.1) (the error term) and at the same time to penalize large (in the nuclear norm sense) values of X. Essentially, the second term reflects our prior knowledge that the solution we seek is low-rank and is forcing the optimization task 6.1.1 to take that into account. This technique in optimization theory is called *regularization* and is encountered in many different scenarios. [BV04] Notice that parameter λ specifies the trade-off between the error term and the regularization term (the nuclear norm term) in the above optimization task. The higher the value of λ, the lower the rank of the recovered solution, but at the cost of increasing the error. Finally, from now on we will refer to the above task as nuclear norm minimization.

We define the functions $f(X) = \frac{1}{2}\|P_\Omega(M) - P_\Omega(X)\|_F^2$ and $g(X) = \lambda\|X\|_*$. Notice that $f(X)$ is a differentiable function, while $g(X)$ is not. However, the problem is a convex one, since it is a sum of convex functions. One possible approach for solving problems of that structure is using the *proximal gradient method* or *proximal forward-backward splitting method* [CP09] [The15]. In order to describe the proximal gradient method we need to define the notion of the *proximal operator*.

Definition 6.1.2 (Proximal operator)**.** *Let $f : \mathbb{R}^n \to \mathbb{R}$. The proximal operator of index δ is an operator*

$$prox_{\delta f} \ : \ \mathbb{R}^n \to \mathbb{R}^n,$$

such that

$$prox_{\delta f}(\boldsymbol{x}) = arg\min_{\boldsymbol{y}\in\mathbb{R}^n} \left\{ f(\boldsymbol{y}) + \frac{1}{2\delta}\|\boldsymbol{x} - \boldsymbol{y}\|_2 \right\}.$$

Notice that the proximal operator is a point in $\mathbb{R}^n$. It can be proven that the previous optimization task is strictly convex (as a sum of a convex and a strictly convex function) and consequently it admits a unique minimum. Also, the proximal operator serves as a generalization of the projection operator, i.e

$$P_C(\boldsymbol{x}) = arg\min_{\boldsymbol{y}\in\mathbb{R}^n} \left\{ i_C(\boldsymbol{y}) + \frac{1}{2}\|\boldsymbol{x} - \boldsymbol{y}\|_2^2 \right\},$$

where $C \subseteq \mathbb{R}^n$ is a nonempty closed convex set and i_C is the indicator function of the set C (returns 1 if $x \in C$ and $+\infty$ otherwise). Now we are ready to introduce the proximal forward-backward splitting method.

Let $f, g : \mathbb{R}^n \to \mathbb{R}$ be two convex functions defined on $\mathbb{R}^n$. Also, assume that f is differentiable, while g is not. Suppose that we have the following optimization task

$$\min_{\boldsymbol{x} \in \mathbb{R}^n} \{f(\boldsymbol{x}) + g(\boldsymbol{x})\}. \tag{6.2}$$

The solutions of the above optimization task are characterized by the following fixed point equation

$$\boldsymbol{x} = prox_{\delta g}(\boldsymbol{x} - \delta \nabla f(\boldsymbol{x})), \forall \delta > 0.$$

Therefore, the proximal gradient method applies the following iteration scheme on the previous optimization task

$$\boldsymbol{x}^{(k+1)} = prox_{\delta_k g}(\boldsymbol{x}^{(k)} - \delta_k \nabla f(\boldsymbol{x}^{(k)})), \tag{6.3}$$

for some appropriately chosen sequence $\{\delta_k\}$.

Essentially, the proximal gradient method is a combination of the proximal method [CP09] and the gradient descent method [The15]. Roughly, it can be shown that the iteration scheme (6.3) converges to a minimum of optimization task (6.2), i.e

$$\boldsymbol{x}^{(k)} \to arg \min_{\boldsymbol{y} \in \mathbb{R}^n} \{f(\boldsymbol{y}) + g(\boldsymbol{y})\},$$

under certain conditions; specifically the differentiable function f must have continuous Lipschitz gradient, i.e

$$\|\nabla f(\boldsymbol{x}) - \nabla f(\boldsymbol{y})\|_2 \leq L\|\boldsymbol{x} - \boldsymbol{y}\|_2, \ \forall \boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^n,$$

for some $L > 0$ and the sequence $\{\delta_k\}$ must be chosen appropriately.

The first step towards adapting the general method of proximal gradient in the nuclear norm minimization problem is to calculate the proximal operator of the nuclear norm.

Proposition 6.1.1 (Proximal operator of the nuclear norm)**.** *Let $M \in \mathbb{R}^{k \times n}$ with SVD $M = U\Sigma V^T$. The proximal operator of the nuclear norm is*

$$prox_{\delta \|\cdot\|_*}(X) = arg \min_{Y \in \mathbb{R}^{k \times n}} \left\{ f(Y) + \frac{1}{2\delta}\|X - Y\|_2 \right\} = D_\delta(X),$$

where $D_\delta(X) = US_\delta(\Sigma)V^T$ and $S_\delta(\Sigma)$ is a diagonal matrix with entries

$$(S_\delta(\Sigma))_{ii} = \begin{cases} \Sigma_{ii} - \delta & \Sigma_{ii} \geq \delta \\ 0 & \Sigma_{ii} < \delta \end{cases}.$$

The operator $D_\delta(X)$ is the soft-thresholding operator for the singular values of X and therefore we can refer to it as *singular value shrinkage operator*.

Also,

$$\begin{aligned} \nabla f(X) &= \nabla \left(\frac{1}{2}\|P_\Omega(M) - P_\Omega(X)\|_F^2 \right) = \\ &= P_\Omega(M) - P_\Omega(X). \end{aligned}$$

As a result, the solutions of the nuclear norm minimization problem are characterized by the following fixed point equation

$$X = D_{\lambda\delta}\left(X - \delta\left[P_\Omega(M) - P_\Omega(X)\right]\right), \forall \delta > 0.$$

Therefore, for our problem the proximal gradient method provides the following iteration scheme

$$X^{(k+1)} = D_{\lambda\delta_k}\left(X^{(k)} - \delta_k\left[P_\Omega(M) - P_\Omega(X^{(k)})\right]\right).$$

In order to simplify the description of the algorithm we use the following notation for the iteration scheme at iteration $i+1$.

$$\begin{cases} Y^{(i+1)} \leftarrow X^{(i)} - \delta_i\left[P_\Omega(M) - P_\Omega(X^{(i)})\right] \\ X^{(i+1)} \leftarrow D_{\lambda\delta_i}(Y^{(i+1)}) = US_{\lambda\delta_i}(\Sigma)V^T \end{cases},$$

where U, Σ, V is the SVD of Y.

We call this algorithm Matrix Completion with Proximal Forward-Backward Splitting (MC-PFBS) and it's pseudocode is provided in Algorithm 1. Notice that we use a fixed step δ.

The complexity per iteration of the above algorithm is $\mathcal{O}(min(k,n)^2 max(k,n))$. Notice that the main load in each iteration is computing the SVD of matrix Y, which is a tractable problem, however for large dimensions this problem becomes quite heavy.

Algorithm 1 Matrix Completion with Proximal Forward-Backward Splitting (MC-PFBS)

Input: $P_\Omega(M) \in \mathbb{R}^{k\times n}$, δ, λ
Output: X
  $X^{(0)}$ with $x_{ij}^{(0)} \sim \mathcal{N}(0,1)$
  while $i < maxiter$ **do**
    $Y^{(i+1)} \leftarrow X^{(i)} - \delta_i\left[P_\Omega(M) - P_\Omega(X^{(i)})\right]$
    $[U, \Sigma, V] \leftarrow SVD(Y^{(i+1)})$
    $X^{(i+1)} \leftarrow US_{\lambda\delta_i}(\Sigma)V^T$
    if $\dfrac{\|X^{(i+1)} - X^{(i)}\|_F}{\|X^{(i)}\|_F} < toler$ **then** break
    end if
  end while

6.1.2 Alternating regularized least squares

One approach to reduce the size of the problem is to factor the matrix $X \in \mathbb{R}^{k\times n}$, with $rank(X) = r$, into two matrices $U \in \mathbb{R}^{k\times r}$ and $V \in \mathbb{R}^{n\times r}$, such that $X = UV^T$. An immediate advantage of this approach is the reduction of the number of variables from $k\cdot n$ to $(k+n)\cdot r$, which is a notable difference in large problems. It can be proven that for the nuclear norm the following result holds.

Proposition 6.1.2. *[DR16] Let $X \in \mathbb{R}^{k\times n}$, then for the nuclear norm of X it holds that*

$$\|X\|_* = \min_{U,V} \frac{1}{2}\left(\|U\|_F^2 + \|V\|_F^2\right) \text{ subject to } X = UV^T.$$

Using the previous proposition we can factorize matrix X into two factors U and V and replace the nuclear norm term in cost function (6.1) with the expression $\frac{1}{2}\left(\|U\|_F^2 + \|V\|_F^2\right)$. This technique is known as *Burer-Monteiro heuristic*, after the authors who proposed this approach for general semidefinite programs. [BM05][BM03] Therefore, we can replace (6.1) with the following cost function

$$\boxed{f_2(U,V) = \frac{1}{2}\|P_\Omega(M) - P_\Omega(UV^T)\|_F^2 + \frac{\lambda}{2}(\|U\|_F^2 + \|V\|_F^2)}, \tag{6.4}$$

where $U \in \mathbb{R}^{k\times L}$ and $V \in \mathbb{R}^{n\times L}$, for some $r \leq L \leq min(k,n)$.

The respective minimization problem is

Definition 6.1.3 (Matrix completion with the Burer-Monteiro heuristic)**.** *Let $M \in \mathbb{R}^{k\times n}$, with $rank(M) = r$. We choose $r \leq L \leq min(k,n)$. Then, we formulate the following optimization task for matrix completion,*

$$\min_{\substack{U\in\mathbb{R}^{n\times L}, V\in\mathbb{R}^{k\times L} \\ s.t\ X=UV^T}} \frac{1}{2}\|P_\Omega(M) - P_\Omega(UV^T)\|_F^2 + \frac{\lambda}{2}(\|U\|_F^2 + \|V\|_F^2).$$

Note that the choice of L should be based on the prior knowledge we have about the nature of the problem we are dealing with. Specifically, the choice of L should be a safe overestimate of the true rank of the matrix. For example, in the Netflix problem the choice of L reflects our belief about the number of essential factors that determine the preference of users for the different movies.

We can solve the above optimization task using the technique known as *alternating minimization*. That means we fix one of the two factor matrices and minimize the cost function with respect to the other factor matrix. Then at each iteration the optimization task that occurs is reduced to a regularized least-squares problem, specifically to a ridge regression problem. Therefore, we call this algorithm *Matrix Completion with Alternating Regularized Least Squares (MC-ARLS)*.

We fix V and we minimize the cost function (6.4) with respect to U. In order to do that we compute the derivative of f_2 with respect to U. After performing all the necessary algebraical manipulations we end up with the following closed-form expression for the jth row of U,

$$\boldsymbol{u}_j = \left(V^TCV + \lambda I\right)^{-1} V^TCY_{(j,\cdot)}^T, \tag{6.5}$$

where $C = diag(P_{(j,\cdot)})$ and $Y = P_\Omega(M)$.

Using the same reasoning, for the jth row of V we have

$$\boldsymbol{v}_j = \left(U^TCU + \lambda I\right)^{-1} U^TCY_{(\cdot,j)}, \tag{6.6}$$

where $C = diag(P_{(\cdot,j)})$ and $Y = P_\Omega(M)$.

The pseudocode of this algorithm is given in algorithm 2. The complexity per iteration of the above algorithm is $\mathcal{O}(mL^2 + max(k,n)L^3)$.

Algorithm 2 Matrix Completion with Alternating Regularized Least Squares (MC-ARLS)

Input: $Y = P_\Omega(M) \in \mathbb{R}^{k\times n}$, P, λ, L
Output: $X = UV^T$
  $U^{(0)} \in \mathbb{R}^{k\times L}$ with $u_{ij}^{(0)} \sim \mathcal{N}(0,1)$
  $V^{(0)} \in \mathbb{R}^{n\times L}$ with $v_{ij}^{(0)} \sim \mathcal{N}(0,1)$
  while $i < maxiter$ **do**
    for $j = 1 : k$ **do**
      $C = diag(P_{(j,\cdot)})$
      $\boldsymbol{u}_j^{(i+1)} \leftarrow \left(V^{(i)T} C V^{(i)} + \lambda I\right)^{-1} V^{(i)T} C Y_{(j,\cdot)}^T$
    end for
    for $j = 1 : n$ **do**
      $C = diag(P_{(\cdot,j)})$
      $\boldsymbol{v}_j^{(i+1)} \leftarrow \left(U^{(i+1)T} C U^{(i+1)} + \lambda I\right)^{-1} U^{(i+1)T} C Y_{(\cdot,j)}$
    end for
    $X^{(i+1)} \leftarrow U^{(i+1)} V^{(i+1)T}$
    if $\dfrac{\|X^{(i+1)} - X^{(i)}\|_F}{\|X^{(i)}\|_F} < toler$ **then** break
    end if
  end while

6.1.3 Alternating iteratively reweighted least squares

In [GRK16] the l_1/l_2 norm for a matrix $X \in \mathbb{R}^{k\times n}$ was introduced, i.e

$$\|X\|_{1,2} = \sum_{i=1}^{n} \|\boldsymbol{x}_{c,i}\|_2,$$

where $\boldsymbol{x}_{c,i}$, $i \in \{1, 2, \ldots, n\}$ are the columns of X. In that paper the l_1/l_2 norm was employed in the design of an online low-rank subspace learning scheme. Next, this norm was incorporated in the optimization task of a low-rank matrix factorization problem [GRK17a] and in a nonnegative matrix factorization problem [GRK17b]. Here we use this norm in the formulation of an optimization task for a matrix completion problem. This approach is contained in the following cost function

$$\boxed{f_3(U,V) = \frac{1}{2}\|P_\Omega(M) - P_\Omega(UV^T)\|_F^2 + \delta \sum_{i=1}^{L} \sqrt{\|\boldsymbol{u}_{c,i}\|_2^2 + \|\boldsymbol{v}_{c,i}\|_2^2}}, \qquad (6.7)$$

where $\boldsymbol{u}_{c,i}$,$\boldsymbol{v}_{c,i}$ denote the columns of matrices $U \in \mathbb{R}^{k\times L}$ and $V \in \mathbb{R}^{n\times L}$ respectively.

The respective minimization problem is the following.

Definition 6.1.4 (Matrix completion with l_1/l_2 norm)**.** *Let $M \in \mathbb{R}^{k\times n}$, with $rank(M) = r$. We choose $r \leq L \leq min(k, n)$. Then, we formulate the following*

optimization task for matrix completion,

$$\min_{\substack{U\in\mathbb{R}^{n\times L},V\in\mathbb{R}^{k\times L}\\ s.t\ X=UV^T}} \frac{1}{2}\|P_\Omega(M)-P_\Omega(UV^T)\|_F^2+\delta\sum_{i=1}^{L}\sqrt{\|\boldsymbol{u}_{c,i}\|_2^2+\|\boldsymbol{v}_{c,i}\|_2^2}. \quad (6.8)$$

In order to unveil the intuitive explanation of the usage of the l_1/l_2 norm in a matrix completion problem, we rewrite the above cost function in the following way

$$f_3(U,V)=\frac{1}{2}\|P_\Omega(M)-P_\Omega(UV^T)\|_F^2+\delta\|Z\|_{1,2}=$$
$$=\frac{1}{2}\|P_\Omega(M)-P_\Omega(UV^T)\|_F^2+\delta\sum_{i=1}^{L}\|\boldsymbol{z}_{c,i}\|_2,$$

where

$$Z=\begin{bmatrix} | & | & \dots & | \\ \boldsymbol{u}_{c,1} & \boldsymbol{u}_{c,2} & \dots & \boldsymbol{u}_{c,L} \\ | & | & \dots & | \\ | & | & \dots & | \\ \boldsymbol{v}_{c,1} & \boldsymbol{v}_{c,2} & \dots & \boldsymbol{v}_{c,L} \\ | & | & \dots & | \end{bmatrix}$$

and $\boldsymbol{z}_{c,i}$ are the columns of Z.

The new regularization term penalizes the l_1/l_2 norm of matrix Z. Hence this optimization task promotes solutions such that the columns of Z have elements with small magnitude. Loosely, the above optimization task tries to shrink as many columns of Z as possible, thus minimizing the rank of the recovered solution . As a result, this optimization task promotes low-rank solutions.

We will apply here the same methodology as in the previous case, i.e alternating minimization. Hence, for the jth row of the matrix U we have the following closed-form expression,

$$\boldsymbol{u}_j=\left(V^TCV+D\right)^{-1}V^TCY_{(j,\cdot)}^T, \quad (6.9)$$

where $C=diag(P_{(j,\cdot)})$, $Y=P_\Omega(M)$ and[1]

$$D=\text{diag}\left(\frac{\delta}{\sqrt{\|\boldsymbol{u}_{c,1}\|_2^2+\|\boldsymbol{v}_{c,1}\|_2^2+\eta^2}},\dots,\frac{\delta}{\sqrt{\|\boldsymbol{u}_{c,L}\|_2^2+\|\boldsymbol{v}_{c,L}\|_2^2+\eta^2}}\right).$$

Using the same reasoning, for the the jth row of V we have

$$\boldsymbol{v}_j=\left(U^TCU+D\right)^{-1}U^TCY_{(\cdot,j)}, \quad (6.10)$$

where $C=diag(P_{(\cdot,j)})$ and Y and D are defined as previously.

Notice that in this algorithm we introduced the term D, which considers the values of the columns of the factor matrices U and V at the previous iteration.

[1]We use a small constant η to guarantee smoothness.

That term changes in every iteration and essentially serves as a weighting factor in the formulation of the newest estimate. Therefore, we call this algorithm Matrix Completion with Alternating Iteratively Reweighted Least Squares (MC-AIRWLS) and it's description is provided in algorithm 3. Finally, the complexity per iteration of MC-AIRWLS is $\mathcal{O}(mL^2 + max(k,n)L^3)$.

Algorithm 3 Matrix Completion with Alternating Iteratively Reweighted Least Squares (MC-AIRWLS)

Input: $Y = P_\Omega(M) \in \mathbb{R}^{k\times n}$, P, δ, L
Output: $X = UV^T$

$U^{(0)} \in \mathbb{R}^{k\times L}$ with $u_{ij}^{(0)} \sim \mathcal{N}(0,1)$
$V^{(0)} \in \mathbb{R}^{n\times L}$ with $v_{ij}^{(0)} \sim \mathcal{N}(0,1)$
$D^{(0)} \in \mathbb{R}^{L\times L}$ with $d_{ii}^{(0)} \sim \mathcal{N}(0,1)$
while $i < maxiter$ **do**
  for $j = 1 : k$ **do**
    $C = diag(P_{(j,\cdot)})$
    $\boldsymbol{u}_j^{(i+1)} \leftarrow \left(V^{(i)T} C V^{(i)} + D^{(i)}\right)^{-1} V^{(i)T} C Y_{(j,\cdot)}^T$
  end for
  for $j = 1 : n$ **do**
    $C = diag(P_{(\cdot,j)})$
    $\boldsymbol{v}_j^{(i+1)} \leftarrow \left(U^{(i+1)T} C U^{(i+1)} + D^{(i)}\right)^{-1} U^{(i+1)T} C Y_{(\cdot,j)}$
  end for
  $D^{(i+1)} \leftarrow \text{diag}\left(\frac{\delta}{\sqrt{\|\boldsymbol{u}_{c,1}^{(i+1)}\|_2^2 + \|\boldsymbol{v}_{c,1}^{(i+1)}\|_2^2 + \eta^2}}, \ldots, \frac{\delta}{\sqrt{\|\boldsymbol{u}_{c,L}^{(i+1)}\|_2^2 + \|\boldsymbol{v}_{c,L}^{(i+1)}\|_2^2 + \eta^2}}\right)$
  $X^{(i+1)} \leftarrow U^{(i+1)} V^{(i+1)T}$
  if $\frac{\|X^{(i+1)} - X^{(i)}\|_F}{\|X^{(i)}\|_F} < toler$ **then** break
  end if
end while

6.1.4 Fast alternating regularized least squares

This algorithm also applies the alternating minimization framework to the cost function (6.4), i.e $f_2(U,V)$. However, at each iteration it computes the minimum of an upper bound of an approximation of $f_2(U,V)$, after fixing one of the two variables. Specifically, we fix one of the two variables of $f_2(U,V)$ and compute an upper bound for the 2nd order Taylor expansion of the respective cost function, using an approximation for the Hessian. Finally, we minimize the cost function that is formulated with respect to the other variable. Note that the technique we use was introduced in [GRK17c] for the cost function (6.7).

Suppose that we want to find the new estimate of U at iteration $i+1$, i.e $U^{(i+1)}$. At this stage we have at our disposal the estimates $U^{(i)}$ and $V^{(i)}$. Therefore, we fix the variable V and compute an upper bound for the 2nd order Taylor expansion of $f_2(U, V^{(i)})$ at $U^{(i)}$, which is

$$\widetilde{f}_2(U|U^{(i)},V^{(i)}) = f_2(U^{(i)},V^{(i)}) + \operatorname{tr}\left\{\left(U-U^{(i)}\right)^T \nabla_U f_2(U^{(i)},V^{(i)})\right\} + \operatorname{tr}\left\{\left(U-U^{(i)}\right)\widetilde{H}(U^{(i)},V^{(i)})\left(U-U^{(i)}\right)^T\right\},$$

where

$$\widetilde{H}(U,V) = V^TV + \lambda I.$$

Roughly, the choice of $\widetilde{H}$ is justified by the fact that the matrix $\widetilde{H} - H$ is positive semidefinite (if we arrange the elements of those matrices in a suitable way), where H is the Hessian of $f_2(U,V^{(i)})$. For more details refer to [GRK17c].

Therefore, we need to find the minimum value of the following optimization problem

$$\min_{U\in\mathbb{R}^{k\times L}} \widetilde{f}_2(U|U^{(i)},V^{(i)}).$$

If we compute the derivative of $\widetilde{f}_2(U|U^{(i)},V^{(i)})$ with respect to U, equate the result to 0 and solve the respective equation with respect to U we get

$$U^{(i+1)} = U^{(i)} + \left\{\left[P_\Omega(M) - P_\Omega(U^{(i)}V^{(i)T})\right]V^{(i)} - \lambda U^{(i)}\right\}\left[V^{(i)T}V^{(i)} + \lambda I\right]^{-1}$$

Using the same reasoning, we fix U and formulate the following optimization problem

$$\min_{V\in\mathbb{R}^{k\times L}} \widetilde{f}_2(V|U^{(i+1)},V^{(i)}),$$

where the cost function $\widetilde{f}_2(V|U^{(i+1)},V^{(i)})$ is an upper bound for the 2nd order Taylor expansion of $f_2(U^{(i+1)},V)$ at $V^{(i)}$ and is defined as

$$\widetilde{f}_2(V|U^{(i+1)},V^{(i)}) = f_2(U^{(i+1)},V^{(i)}) + \operatorname{tr}\left\{\left(V-V^{(i)}\right)^T \nabla_V f_2(U^{(i+1)},V^{(i)})\right\} + \operatorname{tr}\left\{\left(V-V^{(i)}\right)\widetilde{J}(U^{(i+1)},V^{(i)})\left(V-V^{(i)T}\right)\right\},$$

where

$$\widetilde{J}(U,V) = U^TU + \lambda I.$$

If we compute the derivative of $\widetilde{f}_2(V|U^{(i+1)},V^{(i)})$ with respect to V, equate the result to 0 and solve the respective equation with respect to V we get

$$V^{(i+1)} = V^{(i)} + \left\{ \left[P_\Omega(M) - P_\Omega(U^{(i+1)} V^{(i)T}) \right]^T U^{(i+1)} - \lambda V^{(i)} \right\} \left[U^{(i+1)T} U^{(i+1)} + \lambda I \right]^{-1}$$

We call this algorithm Matrix Completion with Fast Alternating Regularized Least Squares (MC-FARLS) and we provide it's pseudocode in algorithm 4. The complexity per iteration of MC-FARLS is $\mathcal{O}(mL + max(k,n)L^2)$.

Algorithm 4 Matrix Completion with Fast Alternating Regularized Least Squares (MC-FARLS)

Input: $P_\Omega(M) \in \mathbb{R}^{k \times n}$, λ, L
Output: $X = UV^T$
  $U^{(0)} \in \mathbb{R}^{k \times L}$ with $u_{ij}^{(0)} \sim \mathcal{N}(0,1)$
  $V^{(0)} \in \mathbb{R}^{n \times L}$ with $v_{ij}^{(0)} \sim \mathcal{N}(0,1)$
  while $i < maxiter$ **do**
    $U^{(i+1)} \leftarrow U^{(i)} + \left\{ \left[P_\Omega(M) - P_\Omega(U^{(i)} V^{(i)T}) \right] V^{(i)} - \lambda U^{(i)} \right\} \left[V^{(i)T} V^{(i)} + \lambda I \right]^{-1}$
    $V^{(i+1)} \leftarrow V^{(i)} + \left\{ \left[P_\Omega(M) - P_\Omega(U^{(i+1)} V^{(i)T}) \right]^T U^{(i+1)} - \lambda V^{(i)} \right\} \left[U^{(i+1)T} U^{(i+1)} + \lambda I \right]^{-1}$
    $X^{(i+1)} \leftarrow U^{(i+1)} V^{(i+1)T}$
    if $\dfrac{\|X^{(i+1)} - X^{(i)}\|_F}{\|X^{(i)}\|_F} < toler$ **then** break
    end if
  end while

6.1.5 Fast alternating iteratively reweighted least squares

This algorithm applies the procedure we described in the previous subsection to the cost function (6.7), i.e $f_3(U,V)$. This approach is contained in [GRK17c]. Therefore, if we fix variable V we get the following minimization problem

$$\min_{U \in \mathbb{R}^{k \times L}} \widetilde{f}_3(U | U^{(i)}, V^{(i)}).$$

The cost function $\widetilde{f}_3(U|U^{(i)}, V^{(i)})$ is an upper bound for the 2nd order Taylor approximation of $f_3(U, V^{(i)})$ at $U^{(i)}$, i.e

$$\widetilde{f}_3(U|U^{(i)}, V^{(i)}) = f_3(U^{(i)}, V^{(i)}) + \operatorname{tr}\left\{ \left(U - U^{(i)} \right)^T \nabla_U f_3(U^{(i)}, V^{(i)}) \right\} + \operatorname{tr}\left\{ \left(U - U^{(i)} \right) \widetilde{H}(U^{(i)}, V^{(i)}) \left(U - U^{(i)T} \right)^T \right\},$$

where

$$\widetilde{H}(U,V) = V^T V + D$$

and

$$D^{(i)} = \text{diag}\left(\frac{\delta}{\sqrt{\|\boldsymbol{u}_1^{(i)}\|_2^2 + \|\boldsymbol{u}_1^{(i)}\|_2^2 + \eta^2}}, \ldots, \frac{\delta}{\sqrt{\|\boldsymbol{u}_L^{(i)}\|_2^2 + \|\boldsymbol{u}_L^{(i)}\|_2^2 + \eta^2}} \right).$$

If we compute the derivative of $\widetilde{f}_3(U|U^{(i)}, V^{(i)})$ with respect to U, equate the result to 0 and solve the respective equation with respect to U we get

$$U^{(i+1)} = U^{(i)} + \left\{ \left[P_\Omega(M) - P_\Omega(U^{(i)}V^{(i)T}) \right] V^{(i)} - U^{(i)}D^{(i)} \right\} \left[V^{(i)T}V^{(i)} + D^{(i)} \right]^{-1}.$$

In a similar way, if we fix U we formulate the following minimization problem

$$\min_{V \in \mathbb{R}^{k \times L}} \widetilde{f}_3(V|U^{(i+1)}, V^{(i)}).$$

The cost function $\widetilde{f}_3(V|U^{(i+1)}, V^{(i)})$ is an upper bound for the 2nd order Taylor approximation of $f_3(U^{(i+1)}, V)$ at $V^{(i)}$, i.e

$$\begin{aligned} \widetilde{f}_3(V|U^{(i+1)}, V^{(i)}) = f_3(U^{(i+1)}, V^{(i)}) &+ \text{tr}\left\{ \left(V - V^{(i)} \right)^T \nabla_V f_3(U^{(i+1)}, V^{(i)}) \right\} \\ &+ \text{tr}\left\{ \left(V - V^{(i)} \right) \widetilde{J}(U^{(i+1)}, V^{(i)}) \left(V - V^{(i)T} \right) \right\}, \end{aligned}$$

where we have that

$$\widetilde{J}(U, V) = U^T U + D.$$

If we compute the derivative of $\widetilde{f}_3(V|U^{(i+1)}, V^{(i)})$ with respect to V, equate the result to 0 and solve with respect to V we get

$$V = V^{(i)} + \left\{ \left[P_\Omega(M) - P_\Omega(U^{(i+1)}V^{(i)T}) \right]^T U^{(i+1)} - V^{(i)}D^{(i)} \right\} \left[U^{(i+1)T}U^{(i+1)} + D^{(i)} \right]^{-1}$$

We call this algorithm Matrix Completion with Fast Alternating Iteratively Reweighted Least Squares (MC-AIRWLS) and we provide it's pseudocode in algorithm 5. The complexity per iteration of MC-FAIRWLS is $\mathcal{O}(mL+max(k,n)L^2)$.

Figure 6.1 contains the complexity per iteration of the 5 algorithms we use.

6.2 Evaluation on synthetic data

In this section we will evaluate the five algorithms we described previously on synthetic data.

Algorithm 5 Matrix Completion with Fast Alternating Iteratively Reweighted Least Squares (MC-AIRWLS)

Input: $P_\Omega(M) \in \mathbb{R}^{k\times n}$, δ, L
Output: $X = UV^T$

$U^{(0)} \in \mathbb{R}^{k\times L}$ with $u_{ij}^{(0)} \sim \mathcal{N}(0,1)$
$V^{(0)} \in \mathbb{R}^{n\times L}$ with $v_{ij}^{(0)} \sim \mathcal{N}(0,1)$
$D^{(0)} \in \mathbb{R}^{L\times L}$ with $d_{ii}^{(0)} \sim \mathcal{N}(0,1)$
while $i < maxiter$ **do**

$$U^{(i+1)} \leftarrow U^{(i)} + \left\{ \left[P_\Omega(M) - P_\Omega(U^{(i)}V^{(i)T})\right] V^{(i)} - U^{(i)}D^{(i)} \right\} \left[V^{(i)T}V^{(i)} + D^{(i)}\right]^{-1}$$

$$V^{(i+1)} \leftarrow V^{(i)} + \left\{ \left[P_\Omega(M) - P_\Omega(U^{(i+1)}V^{(i)T})\right]^T U^{(i+1)} - V^{(i)}D^{(i)} \right\} \left[U^{(i+1)T}U^{(i+1)} + D^{(i)}\right]^{-1}$$

$$D^{(i+1)} \leftarrow \operatorname{diag}\left(\frac{\delta}{\sqrt{\|\boldsymbol{u}_1^{(i+1)}\|_2^2 + \|\boldsymbol{u}_1^{(i+1)}\|_2^2 + \eta^2}}, \ldots, \frac{\delta}{\sqrt{\|\boldsymbol{u}_L^{(i+1)}\|_2^2 + \|\boldsymbol{u}_L^{(i+1)}\|_2^2 + \eta^2}} \right)$$

$X^{(i+1)} \leftarrow U^{(i+1)}V^{(i+1)T}$
if $\dfrac{\|X^{(i+1)} - X^{(i)}\|_F}{\|X^{(i)}\|_F} < toler$ **then** break
end if
end while

algorithm	**Complexity per Iteration (CPI)**
MC-PFBS	$\mathcal{O}(min(k,n)^2 max(k,n))$
MC-ARLS	$\mathcal{O}(mL^2 + max(k,n)L^3)$
MC-AIRWLS	$\mathcal{O}(mL^2 + max(k,n)L^3)$
MC-FARLS	$\mathcal{O}(mL + max(k,n)L^2)$
MC-FAIRWLS	$\mathcal{O}(mL + max(k,n)L^2)$

Figure 6.1: The Complexity per Iteration (CPI) of the 5 algorithms we use.

	S1	S2	S3	S4	S5	S6	S7	S8
$\boldsymbol{k, n}$	$150, 300$	$150, 300$	$150, 300$	$150, 300$	$150, 300$	$150, 300$	$150, 300$	$150, 300$
$\mathbf{p}$	50	50	50	50	75	75	75	75
$\boldsymbol{\sigma}$	10^{-1}	10^{-1}	10^{-2}	10^{-2}	10^{-1}	10^{-1}	10^{-2}	10^{-2}
$\mathbf{r}$	10	20	10	20	10	20	10	20

Figure 6.2: The parameters of the 8 scenarios.

6.2.1 Description of the scenarios

We generate $k \times n$ matrices of (true) rank r as a product of two standard Gaussian random matrices of dimensions $k \times r$ and $r \times n$ respectively. On that matrix we sample uniformly at random m entries (the observed entries) and set the rest to zero. We assume that there is some measurement noise that we incorporate by contaminating the observed entries of the previous matrix with zero-mean Gaussian noise of standard deviation σ. In order to evaluate the performance of the 5 algorithms we develop 8 scenarios by altering the parameters of the randomly generated matrices and the noise level. The parameters we consider are the following :

- $\boldsymbol{k, n}$: the dimensions of the matrix
- $\boldsymbol{r}$: the true rank of the matrix
- $\boldsymbol{\sigma}$: the standard deviation of the zero-mean Gaussian noise that contaminates the observed entries of the synthetic matrix
- $\boldsymbol{p}$: percentage of missing entries, i.e $p = 1 - \dfrac{m}{k \cdot n}$

The 8 scenarios are contained in figure 6.2.1.

We perform the experiments for each scenario in the following way. We fix the algorithm's parameters and run the same experiment 50 times, randomly drawing in each run a different matrix M. The results we produce in the end are averaged over 50 runs.

Another important issue is the stopping criterion we use to terminate the algorithms. We use the following condition,

$$\frac{\|X^{(i+1)} - X^{(i)}\|_F}{\|X^{(i)}\|_F} < toler,$$

where *toler* is the parameter that specifies how small the relative distance between two consecutive estimates must be in order for the algorithm to terminate, i.e the termination tolerance. We set in all experiments and algorithms the value $toler = 10^{-4}$.

6.2.2 Performance measures

We are going to evaluate the algorithms on the following performance measures.

1. Relative error in the Frobenius norm sense

algorithm	parameters	**S1**	**S2**	**S3**	**S4**	**S5**	**S6**	**S7**	**S8**
MC-PFBS	δ	0.4	0.17	2	0.53	0.1	0.21	0.1	0.53
	λ	2.5	6	0.5	1.9	10	4.7	10	1.9
MC-ARLS	λ	3.5	0.1	0.6	0.1	0.3	0.3	0.1	0.3
	L	15	25	15	25	15	25	15	25
MC-AIRWLS	δ	6	6.2	2.5	3.3	2	6.3	1.1	6.1
	L	15	25	15	25	15	25	15	25
MC-FARLS	λ	4.6	0.1	4.1	0.6	0.1	1.1	0.6	1.1
	L	15	25	15	25	15	25	15	25
MC-FAIRWLS	δ	9.1	9.1	4.6	7	5	20.6	5.6	20.6
	L	15	25	15	25	15	25	15	25

Figure 6.3: The values of the parameters we used in the experiments for the 5 algorithms, across the 8 different scenarios.

The relative error in the Frobenius norm sense is

$$\text{relative error} = \frac{\|M - X\|_F}{\|M\|_F},$$

where X is the final estimate produced by the respective algorithm.

2. Rank

   The rank of the final estimate, i.e the rank of the recovered solution.

3. Time

   The time the algorithm needs to terminate and produce it's output (the final estimate).

Note that for every algorithm we return one value for each performance measure in each scenario, which is the average of the respective values of that measure on the 50 runs of the experiment.

6.2.3 Algorithm parameters

All the algorithms we study and evaluate have parameters that specify their behavior and whose values need to be chosen appropriately. In order to find the suitable parameters, we tested for each scenario several different parameter values and picked the optimal ones. The optimality lies, in the majority of cases, in choosing the parameter value that makes the algorithm return the solution with the smallest possible relative error, while establishing that the recovered rank is approximately equal to the true one. However, it is also possible that the parameter values that offer solutions with rank equal to the true one, suffer from high relative errors. In that cases we choose the parameter value that provides the smallest possible relative error. In figure 6.3 we give a table that contains the parameter values we used in our experiments for all five algorithms across the eight scenarios.

p	50			
σ	10^{-1}		10^{-2}	
r	10	20	10	20
scenario	**S1**	**S2**	**S3**	**S4**
MC-PFBS	$3.74 \cdot 10^{-2}(10.02)$	$9.05 \cdot 10^{-2}(20.2)$	$6.36 \cdot 10^{-3}(10)$	$2.97 \cdot 10^{-2}(20.66)$
MC-ARLS	$4.72 \cdot 10^{-2}(10.08)$	$2.07 \cdot 10^{-2}(25)$	$1.05 \cdot 10^{-2}(10.08)$	$7.00 \cdot 10^{-3}(25)$
MC-AIRWLS	$2.42 \cdot 10^{-2}(10)$	$1.95 \cdot 10^{-2}(20.14)$	$5.40 \cdot 10^{-3}(10.22)$	$6.10 \cdot 10^{-3}(20.24)$
MC-FARLS	$5.89 \cdot 10^{-2}(10.42)$	$2.11 \cdot 10^{-2}(25)$	$4.89 \cdot 10^{-2}(10.06)$	$1.08 \cdot 10^{-2}(24.96)$
MC-FAIRWLS	$2.37 \cdot 10^{-2}(10.08)$	$1.94 \cdot 10^{-2}(20.3)$	$4.69 \cdot 10^{-3}(10.08)$	$6.29 \cdot 10^{-3}(20.04)$

Figure 6.4: The relative error and the recovered rank of the five algorithms across scenarios S1-S4.

p	75			
σ	10^{-1}		10^{-2}	
r	10	20	10	20
scenario	**S5**	**S6**	**S7**	**S8**
MC-PFBS	$2.97 \cdot 10^{-1}(10.74)$	$4.66 \cdot 10^{-1}(56.54)$	$2.96 \cdot 10^{-2}(10.62)$	$4.5 \cdot 10^{-1}(64)$
MC-ARLS	$4.06 \cdot 10^{-2}(15)$	$2.71 \cdot 10^{-12}(25)$	$1.44 \cdot 10^{-2}(15)$	$1.96 \cdot 10^{-1}(25)$
MC-AIRWLS	$2.68 \cdot 10^{-2}(10.12)$	$1.21 \cdot 10^{-1}(20.66)$	$7.30 \cdot 10^{-3}(10.14)$	$1.39 \cdot 10^{-1}(20.84)$
MC-FARLS	$4.38 \cdot 10^{-2}(15)$	$2.16 \cdot 10^{-1}(25)$	$3.38 \cdot 10^{-2}(15)$	$2.11 \cdot 10^{-1}(25)$
MC-FAIRWLS	$2.88 \cdot 10^{-2}(10.14)$	$1.31 \cdot 10^{-1}(20.06)$	$1.31 \cdot 10^{-2}(10.2)$	$1.28 \cdot 10^{-1}(20.04)$

Figure 6.5: The relative error and the recovered rank of the five algorithms across scenarios S5-S8.

6.2.4 Results

Figures 6.4 and 6.5 contain the results of the experiments we run, i.e the relative error and the recovered rank. Also, in figure 6.6 the runtime for all 5 algorithms across the 8 scenarios is provided. Finally, in figure 6.7 the relative error in scenario 3, of the five algorithms, with respect to the number of iterations is illustrated.

p	50				75			
σ	10^{-1}		10^{-2}		10^{-1}		10^{-2}	
r	10	20	10	20	10	20	10	20
scenario	**S1**	**S2**	**S3**	**S4**	**S5**	**S6**	**S7**	**S8**
MC-PFBS	2.78	5.43	4.87	5.53	8.06	14.67	8.32	15.05
MC-ARLS	3.18	16.45	6.83	14.54	20.54	133.53	20.16	134.65
MC-AIRWLS	5.20	9.13	4.73	10.55	11.18	94.04	9.28	104.61
MC-FARLS	0.08	0.80	0.10	0.34	1.67	2.96	0.66	3.03
MC-FAIRWLS	0.24	0.43	0.39	0.54	1.06	2.90	0.95	2.93

Figure 6.6: The runtime of the five algorithms across all 8 scenarios.

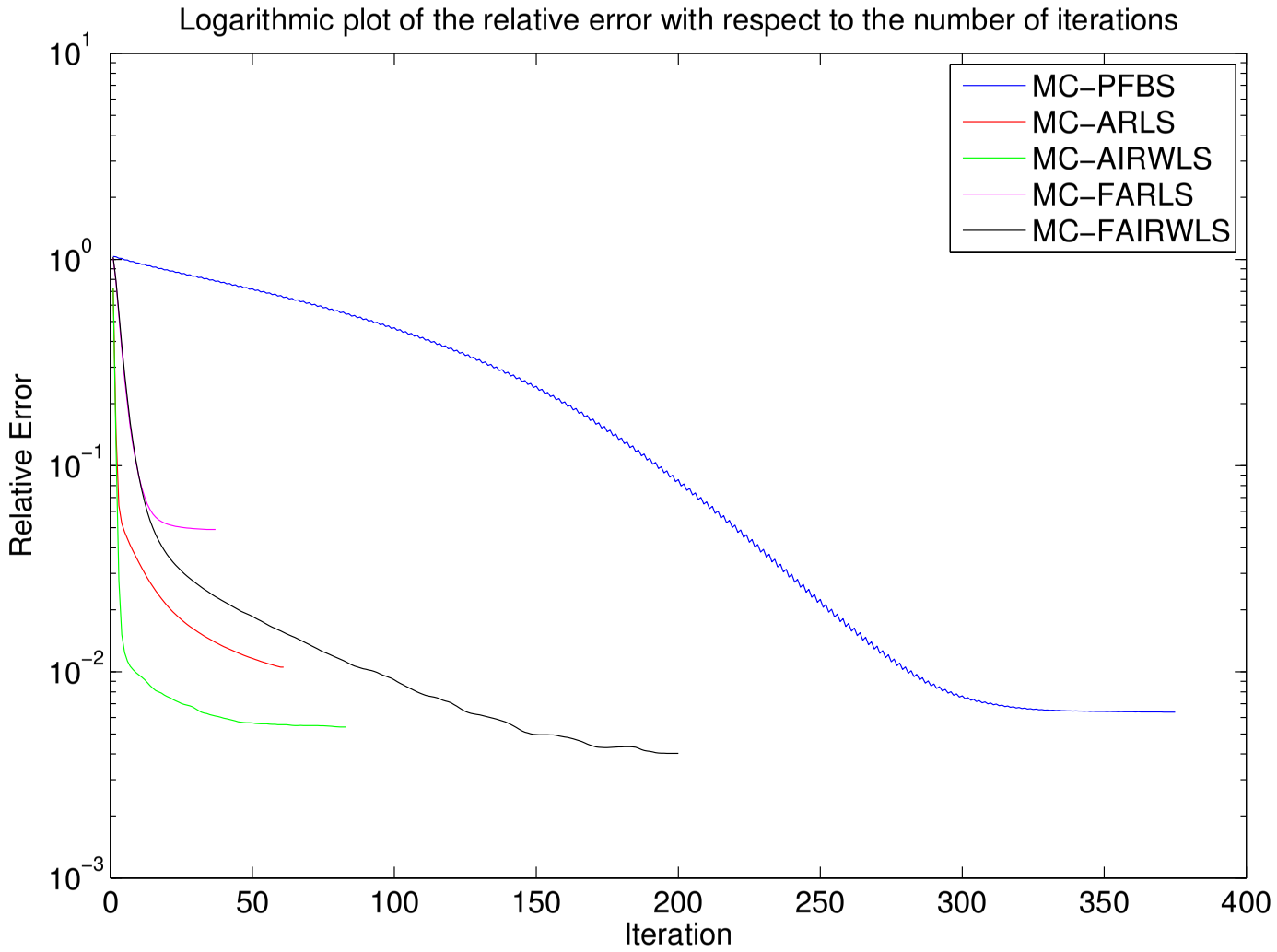

Figure 6.7: Relative error (in logarithmic scale) with respect to the number of iterations (averaged over 50 runs).

6.3 Evaluation on the MovieLens dataset

In this section we will evaluate three algorithms on real data, specifically on the MovieLens dataset. [Mov]

6.3.1 Description of the scenario

In this section we use the MovieLens 100K dataset. This dataset contains 100000 ratings from 943 users on 1682 movies. The ratings are integers in the interval 1-5 and we know that each user has rated at least 20 movies. In those types of datasets we need to split the data into a training and a test set. Instead of doing this on our own we use the splits provided in the dataset's files. Specifically we use "ub.base" as a training set and "ub.test" as a test set. The test set is created by keeping exactly 10 ratings per user from the 100K dataset.

We formulate the rating matrix of the training set by assigning to each user a row of the matrix and to each movie a column. Then, the rating of user i for the movie j is given on the (i, j) entry of the rating matrix. The resulting matrix (dimensions 943×1675) is incomplete and the known entries comprise about 5.7% of the total entries. That is the matrix we seek to complete. We evaluate our result on the ratings contained in the test set, which comprise about 0.6% of the total entries.

We evaluate three algorithms on this dataset, MC-PFBS, MC-FARLS, MC-FAIRWLS. We choose the optimal parameters simply by testing a wide range of parameter values on the given data and selecting the values that make the algorithms return the solutions with the minimum error (note that the notion of error is different in this scenario). Specifically, for MC-PFBS we chose the parameters $\lambda = 3$, $\delta = 1/3$, for MC-FARLS the parameters $\lambda = 7$, $L = 30$ and for MC-FAIRWLS the parameters $\delta = 105$, $L = 30$. Finally, the stopping criterion we use to terminate the algorithms is the same we used on the synthetic data.

6.3.2 Performance measures

Assume that we denote with T the positions of known entries of the incomplete matrix M, which we use to evaluate the 3 algorithms, i.e the test set. Also, let $card(T) = c$. We are going to evaluate the algorithms based on the following performance measures.

1. Root mean square error (RMSE)
   The root mean square error (RMSE) on T is defined as
   $$\text{RMSE} = \frac{\|P_T(M) - P_T(X)\|_F}{\sqrt{c}}.$$

2. Normalized mean average error (NMAE)
   The normalized mean average error (NMAE) on T is defined as
   $$\text{NMAE} = \frac{\sum_{(i,j)\in T} |M_{ij} - X_{ij}|}{4c}.$$

algorithm / performance measure	**NMAE**	**RMSE**	**Rank**	**Time(s)**
MC-PFBS	0.2165	1.1006	180	6988
MC-FARLS	0.2112	1.0704	30.00	260
MC-FAIRWLS	0.1867	0.9529	8	638

Figure 6.8: The NMAE, the RMSE, the recovered rank and the runtime of the algorithms MC-PFBS, MC-FARLS and MC-FAIRWLS.

3. Rank

   The rank of the final estimate, i.e the rank of the recovered solution.

4. Time

   The time it takes for the algorithm to terminate and produce it's output.

6.3.3 Results

In figure 6.8 we provide the results of the experiment, specifically the RMSE, the NMAE, the recovered rank and the runtime of the 3 algorithms. Also, in figures 6.9 and 6.10 we plot the NMAE and the RMSE, respectively, of the 3 algorithms with respect to the number of iterations.

6.4 Remarks and conclusions

The following remarks occur from the parameter selection stage and the experiments we performed on synthetic data.

- The MC-ARLS and the MC-FARLS algorithms in most scenarios (except scenarios 1 and 3) fail to provide a solution with rank equal to the true rank. However, during the parameter selection stage we noticed that it is possible to force the two algorithms to return solutions with rank equal to the true one for appropriate choices of the parameter λ, but with a notable deterioration (with respect to the presented results) in the relative error of the recovered matrix. Therefore, we can say that practically the two algorithms, i.e MC-ARLS and MC-FARLS, cannot recover the true rank.

- The MC-AIRWLS and the MC-FAIRWLS algorithms in all scenarios manage to (approximatelly) recover the true rank. Also, they provide the lowest relative error among the 5 algorithms in all scenarios. Actually, they provide the best performance under the restriction (which we imposed during the parameter selection stage) that the recovered solution has rank approximately equal to the true one. On the contrary, we impose no such restrictions on the MC-ARLS and the MC-FARLS algorithms (except from two scenarios where we can obtain the true rank at an affordable cost with resepct to the deterioration of the relative error). This observation accentuates the superior performance of the MC-AIRWLS and the MC-FAIRWLS algorithms.

- With respect to the time required by the algorithms to provide their estimate, we notice that the MC-FARLS and the MC-FAIRWLS algorithms

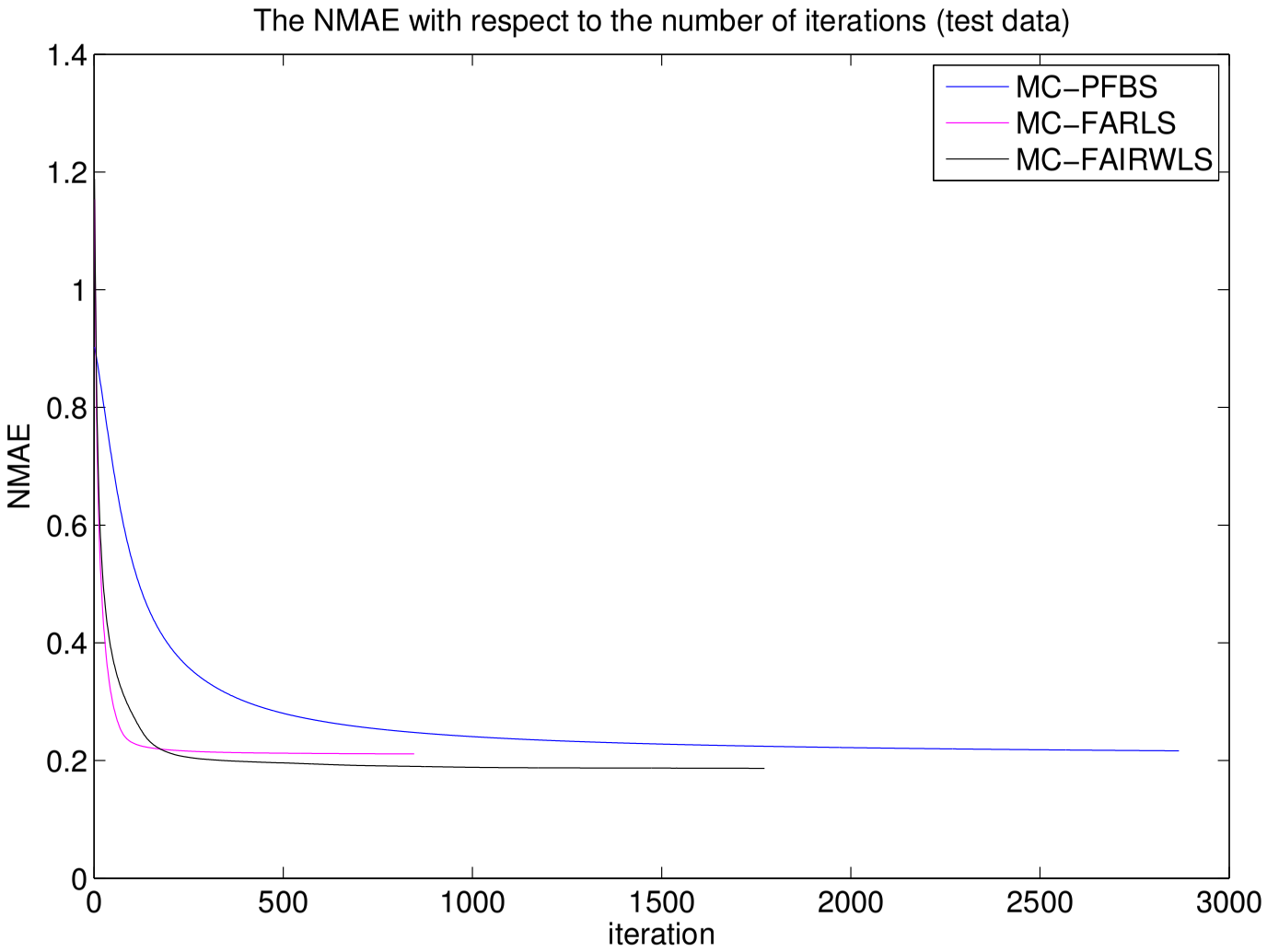

Figure 6.9: The NMAE of MC-PFBS, MC-FARLS and MC-FAIRWLS with respect to the number of iterations.

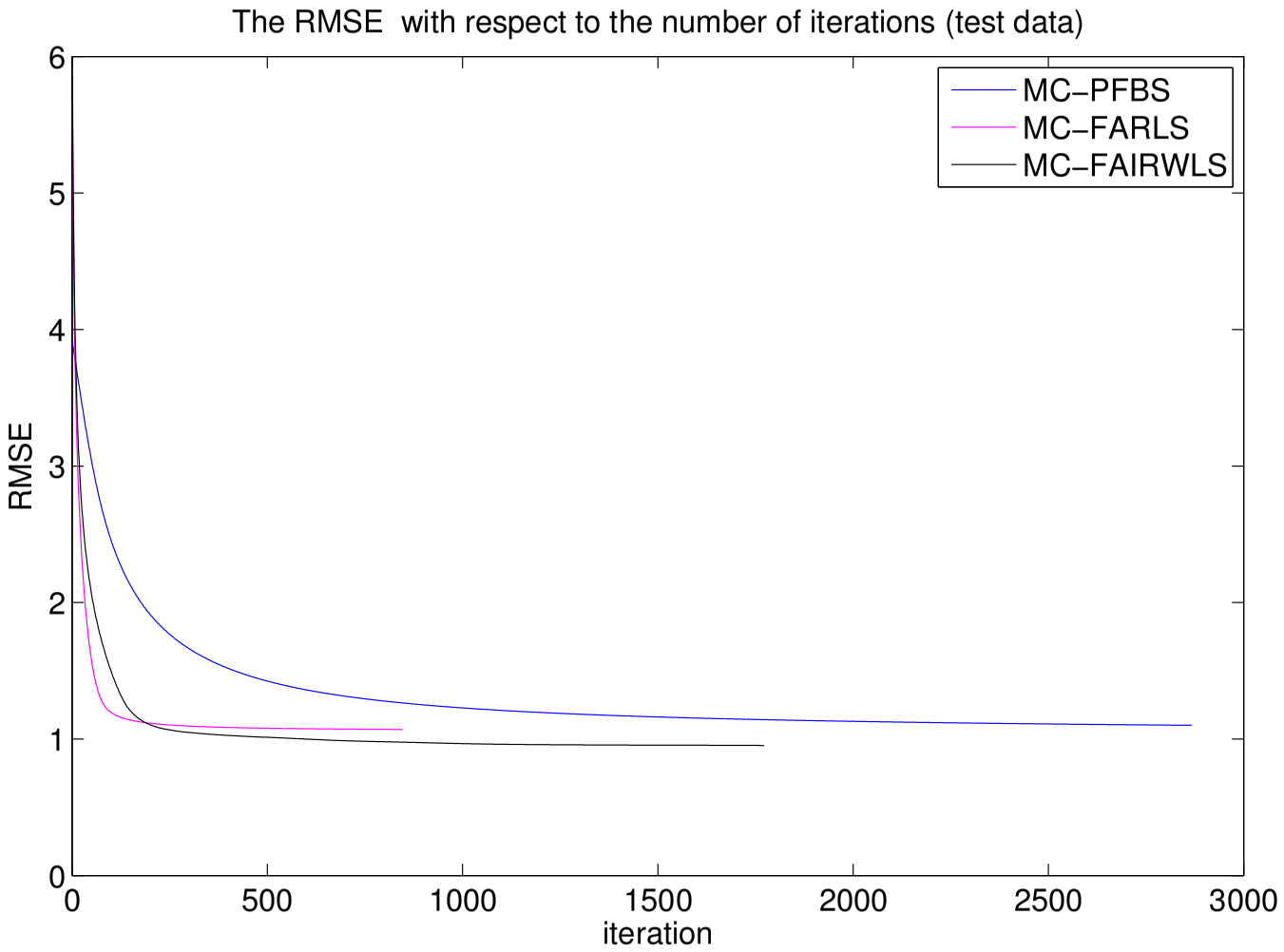

Figure 6.10: The RMSE of MC-PFBS, MC-FARLS and MC-FAIRWLS with respect to the number of iterations.

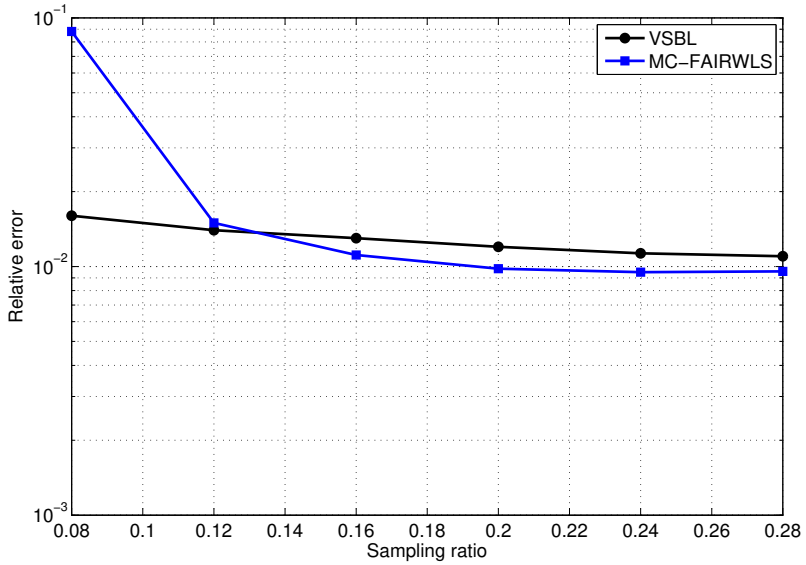

Figure 6.11: The relative error of MC-FAIRWLS and VSBL with respect to the sampling ratio.

are considerably faster than the remaining three algorithms. Especially in the hard scenarios the difference is quite large. That was expected, since the approximation we used in those two algorithms allowed us to compute closed-form expressions for directly updating the factor matrices U and V, while in the MC-ARLS and MC-FARLS algorithms the respective expressions are with resepct to the rows of U and V.

- The MC-AIRWLS and the MC-FAIRWLS algorithms gradually reduce the rank of the estimate during the training stage. This is important, since we can exploit that fact in our implementation to speed up the calculations. On the contrary, we generally (there are some exceptions) do not observe such behaviour in the MC-ARLS and the MC-FARLS algorithms.

We perform a rough comparison with a Bayesian algorithm for matrix completion, the VSBL algorithm.[Bab+12] Indicatively, we perform the experiment depicted in figure 2 (of the respective paper) and using the respective data we compare VSBL with MC-FAIRWLS. In figure 6.11 the relative error with respect to the sampling ratio, of the two algorithms, is illustrated. With respect to the experiment's results, we note that the recovered rank of both algorithms is equal to the true one (in all scenarios), however the runtime of VSBL is higher than the respective runtime of MC-FAIRWLS (which ranges from 20s to 40s) in all scenarios. Also, from figure 6.11 we can see that the relative error of MC-FAIRWLS is sligthly smaller than the respective error of VSBL, for sampling ratios 0.16, 0.2, 0.24, 0.28. However, for highly incomplete matrices (sampling ratios 0.08 and 0.12) VSBL performs better than MC-FAIRWLS and especially for sampling ratio $p = 0.08$ the difference is notable.

The experiments on the MovieLens dataset strengthen some basic observations we made on synthetic data. Moreover, from the figures we notice that the number of iterations that are needed for MC-FAIRWLS to converge is more than two times larger than the respective number for the MC-FARLS algorithm. That is reflected on the runtime of the two algorithms, where MC-FARLS requires less than half of the respective time of MC-FAIRWLS. However, both algorithms work significantly faster than MC-PFBS and thus we can characterize them as time efficient (i.e fast). Finally, we should note that we cannot really evaluate in this scenario the recovered rank, since we are not aware of its true value, however it straightforward to recognize the expected behaviour, as noted in the previous remarks.

In [YDC15] the main algorithms were evaluated against the same dataset and split we use on this thesis, which allows us to make a simple comparison. It is easy to see that MC-FAIRWLS manages to achieve lower RMSE and at the same time to converge using a smaller number of iterations than the respective algorithms.

We can conclude that the algorithms that solve the minimization problem 6.1.4, i.e MC-AIRWLS and MC-FAIRWLS, offer the best performance with respect to the relative error and the recovered rank. If, in addition, we take into account the time needed to compute an output, MC-FAIRWLS offers the superior performance, since it provides solutions with low relative error (in half scenarios it offers the lowest relative error and in the remaining half it provides error close to the lowest one), it manages to recover approximatley the true rank and it works quite fast. The other algorithm whose runtime is comparable to the runtime of MC-FAIRWLS, i.e MC-FARLS, fails to provide comparable performance with respect to the relative error and the recovered rank.

Appendix A

Mathematical Preliminaries

The purpose of this appendix is to collect several lemmas, theorems and definitions that do not fit to the main presentation.

A.1 Linear Algebra

We will start with the fundamental notion of the *norm*.

Definition A.1.1 (Norm)**.** *Let V be a vector space. A norm is a function $\|\cdot\| : V \to \mathbb{R}$ that satisfies the following conditions*

1. $\|\boldsymbol{x}\| \geq 0, \forall \boldsymbol{x} \in V$
2. $\|\boldsymbol{x}\| = 0 \Leftrightarrow \boldsymbol{x} = \mathbf{0}$
3. $\|\lambda \boldsymbol{x}\| = |\lambda| \, \|\boldsymbol{x}\|, \forall \boldsymbol{x} \in V$
4. $\|\boldsymbol{x} + \boldsymbol{y}\| \leq \|\boldsymbol{x}\| + \|\boldsymbol{y}\|, \forall \boldsymbol{x}, \boldsymbol{y} \in V$.

A.1.1 Vectors

Definition A.1.2 (Support)**.** *The support of a vector $\boldsymbol{x} \in \mathbb{R}^n$ is the set of the indices of it's nonzero elements, i.e*

$$supp(\boldsymbol{x}) = \{i \in \{1, \ldots, n\} : x_i \neq 0\}.$$

Vector norms

We will restrict our attention to the vector space $V = \mathbb{R}^n$.

Definition A.1.3 (l_p norms)**.** *The l_p norm of $\boldsymbol{x} \in \mathbb{R}^n$ is defined as*

$$\|\boldsymbol{x}\|_p = \left(\sum_{i=1}^{n} |x_i|^p \right)^{1/p}.$$

The most important l_p norms are the following.

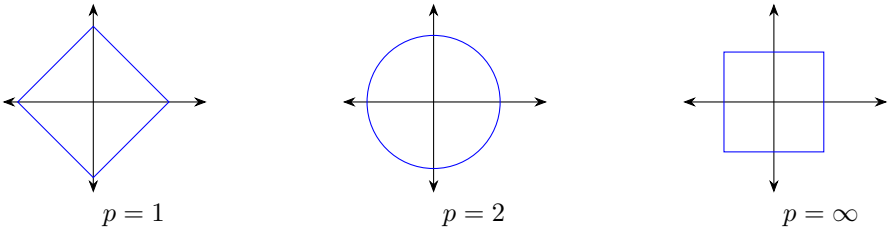

Figure A.1: The unit norm l_p balls for $p = 1, 2, \infty$.

Definition A.1.4 (l_1 norm)**.** *The l_1 norm of $\boldsymbol{x} \in \mathbb{R}^n$ is defined as*

$$\|\boldsymbol{x}\|_1 = \sum_{i=1}^{n} |x_i| .$$

Definition A.1.5 (l_2 norm)**.** *The l_2 norm of $\boldsymbol{x} \in \mathbb{R}^n$ is defined as*

$$\|\boldsymbol{x}\|_2 = \left(\sum_{i=1}^{n} |x_i|^2 \right)^{1/2} .$$

Definition A.1.6 (l_∞ norm)**.** *The l_∞ norm of $\boldsymbol{x} \in \mathbb{R}^n$ is defined as*

$$\|\boldsymbol{x}\|_\infty = \max_{1 \le i \le n} |x_i| .$$

Figure A.1.1 illustrates the unit norm balls for the l_p norms with $p = 1, 2, \infty$.

A.1.2 Matrices

Matrix norms

Now we will provide some norms for matrices. We will define two kind of matrix norms, *Schatten norms* and *operator norms*. The definition of the former follows.

Definition A.1.7 (Schatten norm)**.** *The Schatten p norm of $A \in \mathbb{R}^{k \times n}$ is defined as*

$$\|A\|_p = \left(\sum_{i=1}^{r} |\sigma_i|^p \right)^{1/p} ,$$

where σ_i are the singular values and r the rank of matrix A.

Two notable cases of Schatten norms are the *nuclear norm* and the *Frobenius norm*.

Definition A.1.8 (Nuclear norm)**.** *The nuclear norm of $A \in \mathbb{R}^{k \times n}$ is defined as*

$$\|A\|_* = \sum_{i=1}^{r} |\sigma_i| .$$

Definition A.1.9 (Frobenius norm)**.** *The Frobenius norm of* $A \in \mathbb{R}^{k\times n}$ *is defined as*

$$\|A\|_2 = \left(\sum_{i=1}^{r} |\sigma_i|^2\right)^{1/2} = \left(\sum_{i=1}^{k}\sum_{j=1}^{n} |a_{ij}|^2\right)^{1/2}.$$

Next, we provide the definition of matrix operator norms.

Definition A.1.10 (Matrix operator norms)**.** *The matrix operator norm of* $A \in \mathbb{R}^{k\times n}$ *from* l_p *to* l_q *is defined as*

$$\|A\|_{p\to q} = \sup_{\|\boldsymbol{x}\|_p \le 1} \|A\boldsymbol{x}\|_q.$$

A result that will proven to be useful is contained in the following lemma.

Lemma A.1.1. *[FR13] For the Frobenius and the* $2 \to 2$ *norm the following inequality holds*

$$\|A - B\|_{2\to 2} \le \|A - B\|_F.$$

A.2 Convex geometry

A fundamental notion in convex geometry is the *convex cone*.

Definition A.2.1 (Cone and convex cone)**.** *[FR13] The set* $C \subseteq \mathbb{R}^n$ *is called a cone if*

$$t\boldsymbol{x} \in C,\ \forall \boldsymbol{x} \in C,\ \forall t \ge 0.$$

Furthermore, the set $C \subseteq \mathbb{R}^n$ *is called a convex cone if*

$$t\boldsymbol{x} + s\boldsymbol{y} \in C,\ \forall \boldsymbol{x}, \boldsymbol{y} \in C,\ \forall t, s \ge 0.$$

Essentially a convex cone is a convex set C that is a cone and at the same time. Given a cone C, the *polar cone* of C is defined as follows.

Definition A.2.2 (Polar cone)**.** *[FR13] The polar cone of a cone* C *is*

$$C^\circ = \{\boldsymbol{z} \in \mathbb{R}^n : \langle \boldsymbol{x}, \boldsymbol{z}\rangle \le 0,\ \forall \boldsymbol{x} \in C\}.$$

We are going to use the following lemma in chapter 5.

Lemma A.2.1. *[FR13][Cha+12b] Let* C *be a convex cone,* C° *be it's polar cone and* $\boldsymbol{g} \in \mathbb{R}^n$ *be a vector. Then*

$$\max_{\boldsymbol{z}\in C, \|\boldsymbol{z}\|_2 \le 1} \langle \boldsymbol{g}, \boldsymbol{z}\rangle \le \min_{\boldsymbol{z}\in C^\circ} \|\boldsymbol{g} - \boldsymbol{z}\|_2.$$

A.3 Analysis

Lebesgue's dominated convergence theorem is a classical result in measure theory.

Theorem A.3.1. *[Lebesgue's dominated convergence theorem] [Bre10] Let* $\{f_n\}_{n\in\mathbb{N}}$ *be a sequence of functions such that the following conditions hold*

1. $f_n \in L_1 = \left\{ f : \Omega \to \mathbb{R} \,|\, \int\limits_{\Omega} |f(x)|\, dx < +\infty \right\}$, $\forall n \in \mathbb{N}$,
2. $\lim\limits_{n \to +\infty} f_n(x) = f(x)$ *almost everywhere on* Ω,
3. *There exists a function* $g \in L_1$ *such that* $|f_n(x)| \leq g(x)$, *for all* $n \in \mathbb{N}$, *almost everywhere on* Ω.

   Then, $f \in L_1$ *and*

$$\lim_{n \to +\infty} \int\limits_{\Omega} f_n(x) dx = \int\limits_{\Omega} f(x) dx.$$

A.3.1 Beta and Gamma functions

Beta and Gamma functions will play an important role in this thesis.

Gamma function

The *Gamma function* serves as a continuous analog of the factorial function.[FR13]

Definition A.3.1 (Gamma function)**.** *For* $x > 0$ *the Gamma function is*

$$\Gamma(x) = \int\limits_{0}^{\infty} t^{x-1} e^{-t} dt. \tag{A.1}$$

Lemma A.3.1 (Gamma function properties)**.** *The Gamma function satisfies the following properties*

1. $$\Gamma(x+1) = x\Gamma(x),\ x > 0 \tag{A.2}$$
2. $$\Gamma(n+1) = n!, \tag{A.3}$$

   for all positive integers n.
3. $$\Gamma\left(\frac{1}{2}\right) = \sqrt{\pi} \tag{A.4}$$

Beta function

The definition of *Beta functions* is provided below.[FR13]

Definition A.3.2 (Beta function)**.** *For* $x, y > 0$ *the Beta function is defined as*

$$B(x,y) = \int\limits_{0}^{1} u^{x-1}(1-u)^{y-1} du.$$

Proposition A.3.1 (Beta function property)**.** *For all* $x, y > 0$ *it holds that*

$$B(x,y) = \frac{\Gamma(x)\Gamma(y)}{\Gamma(x+y)}.$$

A.4 Covering numbers

Covering numbers often arise in the context of high dimensional probability theory. Practically the covering number of a set T of a metric space (X, d) is the smallest number of balls of a given radius ρ that cover the whole set. The formal definition follows.

Definition A.4.1 (Covering number)**.** *[FR13] Let (X, d) be a metric space, $T \subseteq X$ and $\rho > 0$. The covering number $\mathcal{N}(T, d, \rho)$ is the smallest number such that*

$$T \subseteq \bigcup_{i=1}^{\mathcal{N}} \mathcal{B}(\boldsymbol{x}_i, \rho),$$

where $\mathcal{B}(\boldsymbol{x}_i, \rho) = \{\boldsymbol{x} \in X : d(\boldsymbol{x}, \boldsymbol{x}_i) \leq \rho\}$,for $\boldsymbol{x}_i \in T$, $i \in \{1, \ldots, \mathcal{N}\}$.

Next, we are going to attain a bound on the covering number of a subset of the unit ball in $\mathbb{R}^n$.

Theorem A.4.1 (Bound for the covering number of a unit ball in $\mathbb{R}^n$)**.** *[FR13] Consider the space $(\mathbb{R}^n, \|\cdot\|)$ and the set $T \subseteq \mathcal{B} = \{\boldsymbol{x} \in \mathbb{R}^n : \|\boldsymbol{x}\|_2 \leq 1\}$. Then we have that*

$$\mathcal{N}(T, \|\cdot\|, \rho) \leq \left(1 + \frac{2}{\rho}\right)^n. \tag{A.5}$$

A.5 Miscellanea

In this section we provide some results, without proof, that do no fit to the previous sections.

Lemma A.5.1 (Upper bound for the combinations formula)**.** *Let $n \geq k > 0$ be some integers. Then,*

$$\binom{n}{k} \leq \left(\frac{en}{k}\right)^k.$$

Stirling's formula is a classical result that provides an approximation of the factorial function. A consequence of Stirling's formula is contained in the following lemma.

Lemma A.5.2 (Stirling's formula)**.** *For all positive integers n we have that*

$$n! \geq \sqrt{2\pi} n^n e^{-n}.$$

Finally, we need the following inequality in chapter 4. For more details on how we can obtain this inequality refer to [FR13].

Lemma A.5.3 (Fenchel's inequality)**.**

$$xy \leq e^x + yln(y) - y, \forall x \in \mathbb{R},\ y > 0.$$

Bibliography

[Ame+14] Dennis Amelunxen et al. "Living on the edge: Phase transitions in convex programs with random data". In: *Information and Inference* (2014), iau005.

[Bab+12] S Derin Babacan et al. "Sparse Bayesian methods for low-rank matrix estimation". In: *IEEE Transactions on Signal Processing* 60.8 (2012), pp. 3964–3977.

[BCW10] Richard G Baraniuk, Volkan Cevher, and Michael B Wakin. "Low-dimensional models for dimensionality reduction and signal recovery: A geometric perspective". In: *Proceedings of the IEEE* 98.6 (2010), pp. 959–971.

[BM03] Samuel Burer and Renato DC Monteiro. "A nonlinear programming algorithm for solving semidefinite programs via low-rank factorization". In: *Mathematical Programming* 95.2 (2003), pp. 329–357.

[BM05] Samuel Burer and Renato DC Monteiro. "Local minima and convergence in low-rank semidefinite programming". In: *Mathematical Programming* 103.3 (2005), pp. 427–444.

[Boc+15] Holger Boche et al. "A Survey of Compressed Sensing". In: *Compressed Sensing and its Applications: MATHEON Workshop 2013*. Ed. by Holger Boche et al. Springer International Publishing, 2015, pp. 1–39.

[Bre10] Haim Brezis. *Functional analysis, Sobolev spaces and partial differential equations*. Springer Science & Business Media, 2010.

[BV04] Stephen Boyd and Lieven Vandenberghe. *Convex optimization*. Cambridge university press, 2004.

[Cha+12a] Djalil Chafaï et al. *Interactions between compressed sensing random matrices and high dimensional geometry*. Société Mathématique de France, 2012.

[Cha+12b] Venkat Chandrasekaran et al. "The convex geometry of linear inverse problems". In: *Foundations of Computational mathematics* 12.6 (2012), pp. 805–849.

[CP09] Patrick L Combettes and Jean-Christophe Pesquet. "Proximal splitting methods in signal processing". In: *arXiv preprint arXiv:0912.3522* (2009).

[CR09] Emmanuel J Candès and Benjamin Recht. "Exact Matrix Completion via Convex Optimization". In: *Foundations of computational mathematics* 9.6 (2009), pp. 717–772.

[CT05] Emmanuel J Candes and Terence Tao. "Decoding by linear programming". In: *IEEE transactions on information theory* 51.12 (2005), pp. 4203–4215.

[CW08] Emmanuel J Candès and Michael B Wakin. "An introduction to compressive sampling". In: *IEEE signal processing magazine* 25.2 (2008), pp. 21–30.

[Don06] David L Donoho. "Compressed sensing". In: *IEEE Transactions on information theory* 52.4 (2006), pp. 1289–1306.

[DR16] Mark A Davenport and Justin Romberg. "An overview of low-rank matrix recovery from incomplete observations". In: *IEEE Journal of Selected Topics in Signal Processing* 10.4 (2016), pp. 608–622.

[FG16] Carlos Fernandez-Granda. *Low-rank models*. Optimization-based Data Analysis, NYU. 2016.

[FR13] Simon Foucart and Holger Rauhut. *A mathematical introduction to compressive sensing*. Vol. 1. 3. Springer, 2013.

[Gor88] Y Gordon. "On Milman's inequality and random subspaces which escape through a mesh in Rn". In: *Lecture Notes in Mathematics* (1988), p. 84.

[GRK16] Paris V Giampouras, Athanasios A Rontogiannis, and Konstantinos D Koutroumbas. "Online low-rank subspace learning from incomplete data using rank revealing l2/l1 regularization". In: *Statistical Signal Processing Workshop (SSP), 2016 IEEE*. IEEE. 2016, pp. 1–5.

[GRK17a] Paris V Giampouras, Athanasios A Rontogiannis, and Konstantinos D Koutroumbas. "l1/l2 regularized non-convex low-rank matrix factorization". In: *Signal Processing with Adaptive Sparse Structured Representations (SPARS) , Lisbon, June 2017* (2017).

[GRK17b] Paris V Giampouras, Athanasios A Rontogiannis, and Konstantinos D Koutroumbas. "Low-rank and Sparse NMF for Joint Endmembers' Number Estimation and Blind Unmixing of Hyperspectral Images". In: *arXiv preprint arXiv:1703.05785* (2017).

[GRK17c] Paris V Giampouras, Athanasios A Rontogiannis, and Konstantinos D Koutroumbas. "Low-rank matrix factorization via l1/l2 norm minimization". In: *(in preparation)* (2017).

[Gro11] D. Gross. "Recovering Low-Rank Matrices From Few Coefficients in Any Basis". In: *IEEE Transactions on Information Theory* 57.3 (2011), pp. 1548–1566.

[Ibm] *What is big data?* `https://www-01.ibm.com/software/data/bigdata/what-is-big-data.html`. [Online; accessed 18-February-2017].

[KC07] Jelena Kovacevic and Amina Chebira. "Life beyond bases: The advent of frames (Part I)". In: *IEEE Signal Processing Magazine* 24.4 (2007), pp. 86–104.

[Mov] *MovieLens dataset*. `https://grouplens.org/datasets/movielens/`. Accessed: 2017-05-01.

[Pri90] Roland Priemer. *Introductory signal processing*. World Scientific Publishing Co Inc, 1990.

[Rec11] Benjamin Recht. "A simpler approach to matrix completion". In: *Journal of Machine Learning Research* 12.Dec (2011), pp. 3413–3430.

[Ste+15] Zachary D Stephens et al. "Big data: astronomical or genomical?" In: *PLoS Biol* 13.7 (2015), e1002195.

[The15] Sergios Theodoridis. *Machine learning: a Bayesian and optimization perspective*. Academic Press, 2015.

[TW10] Joel A Tropp and Stephen J Wright. "Computational methods for sparse solution of linear inverse problems". In: *Proceedings of the IEEE* 98.6 (2010), pp. 948–958.

[YDC15] Alp Yurtsever, Quoc Tran Dinh, and Volkan Cevher. "A universal primal-dual convex optimization framework". In: *Advances in Neural Information Processing Systems*. 2015, pp. 3150–3158.