Learning two layer neural networks by fewest samples

Michael Rauchensteiner

Supervisor: Prof. Massimo Fornasier
Advisor: Prof. Massimo Fornasier
Submission Date: 25.06.2018
I assure the single handed composition of this master’s thesis only supported by declared resources.

München, 25.06.2018

(Michael Rauchensteiner)
Zusammenfassung


Abstract

In this thesis we present a method to identify the weights of a generalized feed-forward neural network with two hidden layers by query samples. The network is represented as a composition of linear combinations of ridge functions. Throughout this work, we assume that the ridge functions are three times continuously differentiable and that the ridge directions of both hidden layers are close to an orthonormal system. In the first part we show that the Hessians of the network, which are evaluated randomly on the unit sphere, concentrate around a matrix space spanned by symmetric rank-one matrices, and that these rank-one matrices are built from the weights of the network. In the second part we show that this matrix space can be approximated by finite differences. Therefore, its construction only relies on function evaluations of the network. The recovery of the near rank-one matrices from the approximative space is done by finding the maximizers of the spectral norm on the intersection between the approximative matrix space with the unit Frobenius sphere. Lastly, we apply the method to a neural network with hyperbolic tangents activations and discuss the numerical results.
Acknowledgments

At this point I want to express my deep gratitude for all the support I received during this work. First of all, my thanks go to my supervisor Massimo Fornasier for all the energy and time he invested into guiding me. His continuous encouragement helped me to overcome many obstacles during this thesis, and his honest commitment to my development means a lot to me. I would also like to thank Johannes Maly for many fruitful discussions and technical insights.

Last but not least, I am much obliged to my family and friends for their support, encouragement and having my back at all times.
# Contents

1 Introduction and preliminaries ......................................................... 1
   1.1 Feed-forward neural networks .................................................. 1
   1.2 Ridge functions ................................................................. 3
   1.3 Our problem and related work .................................................. 4
   1.4 Notation and Preliminaries ..................................................... 5
      1.4.1 Matrix Norms ............................................................... 6
      1.4.2 Orthogonal matrices and the singular value decomposition ...... 7
      1.4.3 Kronecker product and the vectorization of matrices ............ 11

2 Problem specification ........................................................................ 13

3 Concentration of the Hessians around $L$ ....................................... 22
   3.1 Approximating $L$ by $L_{PCA}$ ................................................. 22
   3.2 Uniform distribution on the unit-sphere ..................................... 23
   3.3 Concentration of $L_{PCA}$ around $L$ ......................................... 27

4 Approximation of $L$ by finite differences ..................................... 43
   4.1 Approximating $\nabla f, H[f]$ by finite differences ......................... 43
   4.2 Approximation of $L$ ............................................................. 48
   4.3 Dimension reduction ................................................................ 50

5 Recovering rank-1 matrices from $L$ .............................................. 56
   5.1 Finding a low-rank basis in a matrix subspace ......................... 56
   5.2 Distinguishing the ridge directions .......................................... 58

6 Numerical experiment ...................................................................... 66
   6.1 Optimal initial values ............................................................. 68
      6.1.1 Case $f \in \mathcal{F}^{0.5}_{45}(5,45)$ .................................... 68
      6.1.2 Case $f \in \mathcal{F}^{1}_{45}(5,45)$ .................................... 76
   6.2 Random initial values ............................................................ 77
      6.2.1 Case $f \in \mathcal{F}^{0.5}_{45}(5,45)$ .................................... 78
      6.2.2 Case $f \in \mathcal{F}^{1}_{45}(5,45)$ .................................... 78

Conclusion ......................................................................................... 83
Appendices

A Appendix
Chapter 1

Introduction and preliminaries

Artificial neural networks have undeniably gained immense popularity over the past decade, and their ability to handle unstructured data, such as natural language or images, has made them an important tool in machine learning and artificial intelligence. Moreover, models in machine learning, and in particular neural networks, benefit massively from the increase of data that we produce on a daily basis. Therefore, we can expect them to become even more important in the future. This makes neural networks an interesting topic to study. In particular, since there are many unresolved problems in our theoretical understanding of neural networks, which is, to some extent, in contrast with their huge practical success. One of these problems is the question how much samples are needed to train a neural network. In this thesis we present a method that aims to identify the weights of certain feed-forward neural networks with two hidden layers by evaluations of the network.

As argued in [4], understanding how many samples are needed to identify a network (resp. its parameters) might provide an estimation on the minimal amount of data that is required to train a neural network of the same complexity. We want to start by a short introduction to feed-forward neural networks and their connection to ridge functions.

1.1 Feed-forward neural networks

One of the most known types of neural networks is the so called feed-forward neural network (FFNN). A very basic example for a FFNN would be

$$f(x) = \phi \left( b^T \begin{pmatrix} \sigma(a_1^T x + \vartheta_1) \\ \sigma(a_2^T x + \vartheta_2) \\ \vdots \\ \sigma(a_m^T x + \vartheta_m) \end{pmatrix} + \theta \right),$$

where $a_i \in \mathbb{R}^d, b \in \mathbb{R}^m, \vartheta_i, \theta \in \mathbb{R}$ for all $i = 1, \ldots, m$, and $\phi, \sigma$ are univariate functions. The network takes a vector $x \in \mathbb{R}^d$ and maps it onto $\mathbb{R}$. To better understand the terminology connected with neural networks, it helps to think of $f$ as a composition of
smaller 'computational units'. We can divide the evaluation of \( f \), for a given \( x \in \mathbb{R}^d \), into two steps. First, we calculate

\[
s_1(x) := \begin{pmatrix}
\sigma(a_1^T x + \vartheta_1) \\
\sigma(a_2^T x + \vartheta_2) \\
\vdots \\
\sigma(a_m^T x + \vartheta_m)
\end{pmatrix}.
\]

Second, we evaluate

\[
s_2(s_1(x)) := \phi(b^T s_1(x) + \theta),
\]

based on the result of \( s_1 \). Therefore, \( f \) is given by the composition \( s_2 \circ s_1 \). It is common to refer to \( s_1 \), \( s_2 \) as the layers of the network. We can further divide the layers into smaller units. In fact, to compute \( s_1(x) \) we have to evaluate the expressions \( \sigma(a_1^T x + \vartheta_1), \ldots, \sigma(a_m^T x + \vartheta_m) \), to compute \( s_2(y) \) we have to evaluate \( \phi(b^T y + \theta) \). These small units are the neurons of the network; \( s_1 \) consists of \( m \) neurons, while \( s_2 \) contains only one neuron. The output of a neuron is usually called its activation, the activation of a layer is simply the vector representing the combined activations of its neurons. We can view the input \( x \) as the input layer that just computes the identity for each coordinate of \( x \). The last layer, given by \( s_2 \), is called the output layer, since it yields the final output of the network. Every layer between input and output layer is called hidden layer, in our case this is only \( s_1 \). By this terminology, \( f \) is a feed-forward neural network with one hidden layer. Figure 1.1.1 illustrates the structure of a network \( f : \mathbb{R}^4 \to \mathbb{R} \) with \( m = 5 \) hidden neurons.

The functions \( \sigma, \phi \) are the activation functions of the network, and typically are non-linear, non-decreasing functions. Good examples are:
• binary step function: \( x \mapsto \begin{cases} 1, & \text{for } x \geq 0 \\ 0, & \text{for } x < 0 \end{cases} \)

• rectifier function: \( x \mapsto \max(x, 0) \)

• logistic function: \( x \mapsto \frac{1}{1+\exp(-x)} \)

• hyperbolic tangents: \( x \mapsto \tanh(x) \).

A FFNN with \( \ell > 1 \) hidden layers is simply an extension of (1.1.1) to a composition of more layers \( f = s_\ell \circ s_{\ell-1} \circ \cdots \circ s_1 \). The size of the output/input layer are equivalent to the size of the output/input of the function \( f \). The remaining structure of the network is usually somehow given, this includes the activation functions \( \sigma, \phi \), the number of hidden layers and the number of neurons in each hidden layer. We can think of \( f \) as a model that only depends on the parameters \( a_1, \ldots, a_m, b, \theta, \vartheta_1, \ldots, \vartheta_m \). These parameters are called the weights \((a_1, \ldots, a_m, b)\) and the thresholds \((\theta, \vartheta_1, \ldots, \vartheta_m)\) of the network.

As we have seen, neurons are the core building block of feed-forward neural networks and can mathematically be described as a function \( x \mapsto \sigma(a^T x + \theta) \). This connects neural networks to a class of multivariate functions, which is introduced in the upcoming section. For an extensive introduction into the theory of neural networks, in particular feed-forward neural networks, we refer the interested reader to \[8\].

### 1.2 Ridge functions

Ridge functions are functions of several variables that are constant along parallel hyper-planes in their domain. They represent a simple class of multivariate functions \( f : \mathbb{R}^d \to \mathbb{R} \), and usually take the form

\[
 f(x) = g(a^T x),
\]

for some vector \( a \in \mathbb{R}^d, a \neq 0 \), and a scalar function \( g : \mathbb{R} \to \mathbb{R} \). The vector \( a \) is frequently called the ridge direction of \( f \). Note that, by this definition, the neurons of a feed-forward network can be expressed as ridge functions. One particular application of ridge functions is to approximate a complicated function on \( \mathbb{R}^d \) by a sum of ridge functions

\[
 f(x) = \sum_{i=1}^{m} g_i(a_i^T x).
\]

Approximating functions in many variables often suffers from the so called *curse of dimensionality*. In this case, the cost of an optimal algorithm increases exponentially with the dimensionality (e.g. number of variables) of the problem. Functions with a representation (1.2.1) form a class, which under certain conditions allows to break the curse of dimensionality. One particular approach for functions of the type (1.2.1) is to detect its
active subspace, followed by a reduction of the complexity of the problem to the dimension of the active subspace. Here, active subspace refers to the minimal subspace $A \subset \mathbb{R}^d$ that captures the variability of $f$. Differently put, if $x, x_\perp \in \mathbb{R}^d, x_\perp \not\in A$, then

$$f(x + x_\perp) = f(x).$$

In the matter of sums of ridge functions (1.2.1), it is straightforward to see that the active subspace lies in the span of the ridge directions, i.e.

$$A \subset \text{span}\{a_1, \ldots, a_m\},$$

with equality if $g_1, \ldots, g_m$ are not constant.

An example, where this concept is applied, is the recent work [4] by Fornasier, Vybíral and Daubechies. The authors address the uniform approximation of sums of ridge functions

$$f: \mathbb{R}^d \to \mathbb{R}, \quad f(x) = \sum_{i=1}^{m} g_i(a_i^T x), \quad (1.2.2)$$

where $a_1, \ldots, a_m$ are nearly orthonormal and the functions $g_1, \ldots, g_m$ are three times continuously differentiable. One crucial step in their work is the reduction of $f$ to a representation $\tilde{f}: \mathbb{R}^m \to \mathbb{R}$. This is done by approximating the active subspace of $f$. First, use the fact that $\nabla f(x) \in \text{span}\{a_1, \ldots, a_m\}$ to approximate $\text{span}\{a_1, \ldots, a_m\}$ by evaluating the finite-difference approximation of the gradients randomly on the sphere in $\mathbb{R}^d$. Second, take any orthonormal basis $w_1, \ldots, w_m$ of the resulting space and set $W = (w_1 | \ldots | w_m) \in \mathbb{R}^{d \times m}$. Lastly, substitute $f: \mathbb{R}^d \to \mathbb{R}$ by $\tilde{f}: \mathbb{R}^m \to \mathbb{R}$, where $\tilde{f}(y) = f(W y)$. The function $\tilde{f}$ is then given by a sum of ridge functions in $\mathbb{R}^m$:

$$\tilde{f}: \mathbb{R}^m \to \mathbb{R}, \quad \tilde{f}(y) = \sum_{i=1}^{m} g_i(\alpha_i^T y),$$

where $\alpha_i = W^T a_i$. The authors show that, under their setting and by accepting a small error, identifying $\tilde{f}$ leads to an identification of $f$. For the exact arguments we refer to [4, Section 2]. Additionally, the same technique is used in Section 4.3 of this thesis.

In summary, we gave two reasons that make ridge functions and combination of ridge function interesting:

1. Their role as building blocks of functions in many variables, in particular neural networks.
2. The ability to exploit their structure, which makes them easier to handle from a numerical point of view.

### 1.3 Our problem and related work

Usually, learning a neural network means to fit its weights and thresholds to a given dataset. Whereas in our case, 'learning' describes the identification of a given neural
network. Suppose the network is given as a function $f$. Then, we try to identify the parameters of the network by evaluations of $f$. As argued in [4], neural networks with one hidden layer and one output node, which does not compute a non-linearity, can be expressed by sums of ridge functions. Suppose $m \leq d$, $\phi(t) = t$, $\theta = 0$, then (1.1.1) can be expressed as

$$f(x) = b^T \left( \begin{array}{c} \sigma(a_1^T x + \vartheta_1) \\ \sigma(a_2^T x + \vartheta_1) \\ \vdots \\ \sigma(a_m^T x + \vartheta_1) \end{array} \right) = \sum_{i=1}^{m} b_i \sigma(a_i^T x + \vartheta_i),$$

which is of the form (1.2.1). Hence, sums of ridge functions represent certain shallow feed-forward neural networks. In [4] the directions of $\sum_{i=1}^{m} g_i(a_i^T x)$ are identified by recovering the rank-1 matrices $a_1 a_1^T, \ldots, a_m a_m^T$ as the elements that span the space $\text{span}\{a_1 a_1^T, \ldots, a_m a_m^T\}$.

This space is approximated by the finite difference approximation of the Hessians of $f$ on random points of the unit sphere; let us denote this approximative space with $\tilde{L}$. The approximations of $a_1 a_1^T, \ldots, a_m a_m^T$ are then identified as the elements with maximal spectral norm on the intersection of $\tilde{L}$ with the Frobenius unit sphere.

Our primary motivation is to extend their method of recovering sums of ridge functions to networks with two hidden layers. This network topology can also be represented a composition of ridge functions, given by

$$f(x) = \sum_{\ell=1}^{m_1} h_{\ell} (b_{\ell}^T \left( \begin{array}{c} g_1(a_1^T x) \\ g_2(a_2^T x) \\ \vdots \\ g_m(a_m^T x) \end{array} \right)).$$

We devoted the next chapter to a precise description of our problem, which is concluded by an overview over the remaining structure of the thesis. Beforehand, we need to introduce some notation.

### 1.4 Notation and Preliminaries

If not stated otherwise, matrices and vectors will always be over the real numbers $\mathbb{R}$. We will denote matrices with capital letters and denote their elements by the corresponding small letter. For example, let $A \in \mathbb{R}^{d \times m}$ be a matrix. Then the elements of $A$ are denoted by $a_{ki}$ where $k \in \{1, \ldots, d\}, i \in \{1, \ldots, m\}$. In the same context we address the $i^{th}$ column of $A$ by $a_i$ and the $k^{th}$ row as $a_{ki}$. We write $A^T$ for the transposed matrix of $A$ and extend this notation to vectors. For vectors $v, w$ we will denote the matrix $vw^T$ by the tensor product $v \otimes w$, in the symmetric case $v \otimes v$ we often write $v^{\otimes 2}$. In
many cases, the space spanned by the columns of a matrix $A$ will be denoted by $\text{range}(A)$.

For vectors $v, w \in \mathbb{R}^d$ we denote

$$
\|v\|_p = \begin{cases} 
\left(\sum_{i=1}^{d} v_i^p\right)^{\frac{1}{p}} & \text{for } p \geq 1 \\
\max_i |v_i| & \text{for } p = \infty,
\end{cases}
$$

and primarily use the Euclidean norm $\|v\|_2$, the maximum norm $\|v\|_\infty$ and $\|v\|_1$. Further, we denote by

$$
\langle v, w \rangle = \sum_{i=1}^{d} v_i w_i
$$

the Euclidean inner product. We assume the reader is familiar with these norms and their basic properties like the triangle inequality or the Cauchy-Schwarz inequality.

### 1.4.1 Matrix Norms

**Definition 1** (Frobenius Norm). Let $A$ be a $d \times m$ real matrix, then the Frobenius norm $\|\cdot\|_F$ of $A$ is defined as

$$
\|A\|_F := \sqrt{\sum_{i=1}^{d} \sum_{j=1}^{m} a_{ij}^2}.
$$

**Definition 2** (Trace). The trace of a square matrix $A \in \mathbb{R}^{d \times d}$ is given by the sum of its diagonal elements

$$
\text{Tr}(A) := \sum_{i=1}^{d} a_{ii}.
$$

(1.4.1)

It is well known that the trace of a product of $n$ matrices $A_1, \ldots, A_n$ is invariant under cyclic permutations if the products are well-defined, i.e.

$$
\text{Tr}(A_1 \ldots A_n) = \text{Tr}(A_2 \ldots A_n A_1) = \text{Tr}(A_n A_1 \ldots A_{n-1}).
$$

(1.4.2)

The Frobenius norm is induced by the Frobenius inner product:

$$
\langle A, B \rangle_F := \text{Tr}(AB^T),
$$

and we have

$$
\|M\|_F = \sqrt{\langle M, M \rangle} = \text{Tr}(MM^T).
$$
**Definition 3 (Spectral Norm).** Let $M$ be a $m \times n$ real matrix, then the spectral norm $\|\cdot\| : \mathbb{R}^{m \times n} \to \mathbb{R}$ is defined as

$$\|M\| := \sup_{x \in \mathbb{R}^n, \|x\|_2 = 1} \|Mx\|_2,$$

where $\|\cdot\|_2$ is the Euclidean norm for vectors.

Since the Frobenius inner product will be our only inner product that operates on matrices, we will drop the $F$ in the notation. So, if not stated otherwise, $\langle \cdot, \cdot \rangle = \langle \cdot, \cdot \rangle_F$ for matrices.

We say two matrices $M, \tilde{M} \in \mathbb{R}^{m \times n}$ are orthogonal to each other, i.e. $M \perp \tilde{M}$, if they are orthogonal with respect to the Frobenius inner product

$$M \perp \tilde{M} \iff \langle M, \tilde{M} \rangle = 0.$$

The spectral and the Frobenius norm are submultiplicative, i.e. $\|AB\| \leq \|A\|\|B\|$ and $\|AB\|_F \leq \|A\|_F\|B\|_F$. This property can also be mixed to get better estimates.

**Lemma 1.** [6, Lemma 2.10] For matrices $A, B$, such that $AB$ is defined, we have

$$\|AB\|_F \leq \|A\|\|B\|_F \text{ and } \|AB\|_F \leq \|A\|_F\|B\|_F.$$

### 1.4.2 Orthogonal matrices and the singular value decomposition

**Definition 4 (Orthogonal and unitary matrices).** 1. Let $A \in \mathbb{R}^{d \times m}$ with $d \geq m$. We say $A$ is orthogonal, if the columns of $A$ are orthonormal. This is equivalent to

$$A^T A = I_m.$$

2. Let $U \in \mathbb{R}^{d \times d}$ be a square matrix. We say $U$ is unitary, if the columns and the rows of $U$ are orthonormal, i.e.

$$U^T U = UU^T = I_d.$$

**Remark.** This is a slight deviation from the common definition of orthogonality for real matrices. Often, a real matrix with orthonormal columns is called semi-orthogonal or column-orthogonal, and orthogonal matrices are equivalent to our definition of unitary matrices. The reason for our definition is that we will primarily deal with matrices which are rectangular.

The Frobenius norm and the spectral norm are invariant under unitary transformations. This extends to orthogonal rectangular matrices in the following way: Suppose $A \in \mathbb{R}^{d \times m}$ is orthogonal and $M \in \mathbb{R}^{m \times k}$, then

$$\|AM\|_F = \sqrt{\langle AM, AM \rangle} = \sqrt{\text{Tr} (AMM^T A^T)}$$

$$= \sqrt{\text{Tr} (AMM^T A^T)} = \sqrt{\text{Tr} (A^T A M M^T)}$$

$$= \sqrt{\text{Tr} (I_m M M^T)} = \sqrt{\text{Tr} (M M^T)} = \|M\|_F.$$
by the cyclic invariance of the trace-operator. The same equality holds for the spectral norm, i.e. \( \|AM\| = \|M\| \), which follows by

\[
\|AM\| = \max_{x \in \mathbb{R}^k, \|x\|_2 = 1} \|AMx\|_2 = \max_{x \in \mathbb{R}^k, \|x\|_2 = 1} \sqrt{\langle AMx, AMx \rangle} = \max_{x \in \mathbb{R}^k, \|x\|_2 = 1} \sqrt{\langle Mx, Mx \rangle} = \|M\|.
\]

**Definition 5** (Singular Value Decomposition). Let \( A \in \mathbb{R}^{m \times n} \), then we call

\[
A = U \Sigma V^T,
\]

the singular value decomposition (SVD) of \( A \), if \( U \in \mathbb{R}^{m \times m} \), \( V \in \mathbb{R}^{n \times n} \) are unitary matrices, and \( \Sigma \in \mathbb{R}^{m \times n} \) is a block-diagonal matrix with diagonal elements \( \Sigma_{ii} \geq \ldots \geq \Sigma_{ii} \geq 0 \) for \( i = \min(m, n) \). The diagonal elements of \( \Sigma \) are called singular values of \( A \) and are denoted by \( \sigma_1(A), \ldots, \sigma_{\min(m,n)}(A) \). We will refer to the columns of \( U \), \( V \) as the left and right singular vectors.

The singular value decomposition is an important tool in matrix theory. Some of its properties will be used throughout the next sections.

**Lemma 2.** Let \( A \in \mathbb{R}^{d \times m} \):

1. There exist matrices \( U, V, \Sigma \) such that \( U \Sigma V^T \) is the singular value decomposition of \( A \).

2. Let \( \sigma(A) \) be the vector of singular values of \( A \), then \( \|A\|_F = \|\sigma(A)\|_2 \).

3. The spectral norm of \( A \) is given by its largest singular value: \( \|A\| = \sigma_1(A) \).

4. Let \( k \leq \text{rank}(A) \leq \min(m, n) \) and denote by \( \tilde{\Sigma} \) the matrix that is obtained by setting all but the first \( k \) singular values in \( \Sigma \) to 0. Then \( A_k := U \tilde{\Sigma} V^T \) is the best rank-\( k \) approximation of \( A \) in terms of the Frobenius norm, i.e.

\[
A_k = \arg \min_{M \in \mathbb{R}^{m \times n} \atop \text{rank}(M) = k} \|A - M\|_F.
\]

Proofs to all these properties and a more in depth introduction to the SVD can be found in [6, Section 2.5.3]. Note that one immediate consequence of Lemma 2 is

\[
\|A\|_F = \|\sigma(A)\|_2 \geq \sigma_1(A) = \|A\|.
\]

One important concept in this thesis is quasi-orthogonality. We call a set of vectors quasi-orthogonal if they are close to an orthonormal system of the same size.
Definition 6 (Quasi-orthogonality). Let \( a_1, \ldots, a_m \in \mathbb{R}^d \) such that \( \|a_i\|_2 \leq 1 \) for all \( i = 1, \ldots, m \). Further define

\[
S(a_1, \ldots, a_m) := \inf \left\{ \left( \sum_{i=1}^m \|a_i - \bar{a}_i\|_2^2 \right)^{\frac{1}{2}} \middle| \bar{a}_1, \ldots, \bar{a}_m \in \mathbb{R}^d \text{ orthonormal} \right\}. \tag{1.4.3}
\]

We say \( a_1, \ldots, a_m \) are \( \epsilon \)-quasi-orthonormal (or nearly orthogonal) if there exists an \( \epsilon > 0 \) such that \( S(a_1, \ldots, a_m) \leq \epsilon \).

Remark. By this definition any set of vectors is quasi-orthogonal if their Euclidean norm is bounded by 1 (simply choose \( \epsilon \) sufficiently large). This is clearly not the intuition behind the term. Orthogonality is a binary concept, two vectors are either orthogonal or they are not orthogonal. There is no such distinction for quasi-orthogonality. Hence, the term only gives information about a set of vectors \( a_1, \ldots, a_m \) if we state an \( \epsilon \) such that \( S(a_1, \ldots, a_m) \leq \epsilon \). Nevertheless, outside of proofs we sometimes refer to a set of vectors as quasi-orthogonal, without stating an explicit \( \epsilon \). In this case we imply that there exists an \( \epsilon \) that is reasonably small.

We will often identify a set of vectors as columns of a matrix. The previous terminology can be extended to matrices in the following manner.

Lemma 3. Let \( A = (a_1 \ldots | a_m) \in \mathbb{R}^{d \times m} \) with \( \|a_i\|_2 \leq 1 \). If the columns of \( A \) are quasi-orthogonal with \( \epsilon \), then there is an orthogonal matrix \( \bar{A} \) with the same dimension, such that

\[
\|A - \bar{A}\|_F \leq \epsilon.
\]

Proof. Follows directly by the fact that

\[
\|A - \bar{A}\|_F = \left( \sum_{i=1}^m \|a_i - \bar{a}_i\|_2^2 \right)^{\frac{1}{2}}
\]

for the matrices \( A = (a_1 \ldots | a_m), \bar{A} = (\bar{a}_1 \ldots | \bar{a}_m) \), where \( \bar{a}_1, \ldots, \bar{a}_m \) are orthonormal and

\[
\left( \sum_{i=1}^m \|a_i - \bar{a}_i\|_2^2 \right)^{\frac{1}{2}} < \epsilon.
\]

The existence of the vectors \( \bar{a}_1, \ldots, \bar{a}_m \) is guaranteed by Definition 6. \( \square \)

As a result, we can extend our definition of quasi-orthogonality to matrices if their columns are unit vectors. In the context of the previous lemma, we often refer to the matrix \( \bar{A} \) as the optimal orthogonal matrix of \( A \). In fact, quasi-orthogonality is related to the singular value decomposition in the following way.
**Lemma 4.** \[4, \text{Theorem } 20\] Let \( A \) be a \( d \times m \) matrix with columns of maximal length 1 and \( d \geq m \). Then the columns of \( A \) are quasi-orthonormal with

\[
S(a_1, \ldots, a_m) = \left( \sum_{k=1}^{m} (\sigma_k(A) - 1)^2 \right)^{\frac{1}{2}}.
\]

Further, let \( U\Sigma V^T \) be the singular value decomposition of \( A \) and \( U = (u_1|\ldots|u_d) \). Then the optimal orthogonal matrix w.r.t. \( A \) is given by

\[
\bar{A} = (u_1|\ldots|u_m)V^T.
\]

Throughout the following sections we will make use of the following properties of quasi-orthogonality.

**Lemma 5.** Let \( A \in \mathbb{R}^{d \times m}, B \in \mathbb{R}^{m \times m_1} \) be quasi-orthogonal with \( \epsilon_A, \epsilon_B \in [0, 1] \), and denote by \( \bar{A}, \bar{B} \) their optimal orthogonal matrices. Then the following properties hold:

1. 

\[
(1 - \epsilon_A) \leq \|A\| \leq (1 + \epsilon_A).
\]

2. 

\[
(1 - \epsilon_A)^2 \|S\|_F \leq \|ASA^T\|_F \leq (1 + \epsilon_A)^2 \|S\|_F \tag{1.4.4}
\]

for any matrix \( S \in \mathbb{R}^{m \times m} \).

3. Assume \( S \) is a diagonal matrix, then we can extend 2. to

\[
(1 - 2\epsilon_A) \|S\|_F \leq \|ASA^T\|_F \leq (1 + 2\epsilon_A) \|S\|_F \tag{1.4.5}
\]

4. The product \( AB \) is quasi-orthogonal with \( \epsilon \leq \epsilon_A + \epsilon_B + \epsilon_A \epsilon_B \). Additionally \( \bar{A} \bar{B} \) is orthogonal and

\[
\|AB - \bar{A} \bar{B}\|_F \leq \epsilon_A + \epsilon_B + \epsilon_A \epsilon_B.
\]

**Proof of Lemma 5.** Let \( A \in \mathbb{R}^{d \times m}, B \in \mathbb{R}^{m \times m_1} \) be a quasi-orthogonal with \( \epsilon_A, \epsilon_B \in [0, 1] \), and denote by \( \bar{A}, \bar{B} \) their optimal orthogonal matrices.

1. 

\[
\|A\| \leq \|A + \bar{A} - \bar{A}\| \leq \|\bar{A}\| + \|A - \bar{A}\| = 1 + \epsilon_A,
\]

and by the inverse triangle inequality

\[
\|A\| = \|A - \bar{A} + \bar{A}\| \geq \|\bar{A}\| - \|A - \bar{A}\| = 1 - \epsilon_A.
\]
2. Follows directly by 1. and Lemma 1:

\[ \|ASA^T\|_F \leq \|A\|\|SA^T\|_F \leq \|A\|^2\|S\|_F. \]

3. Denote the columns of \(A\) by \(a_1, \ldots, a_m\), then

\[ \|ASA^T\|_F = \left\| \sum_{i=1}^{m} s_{ii} a_i \otimes a_i \right\|_F, \]

the rest follows by [4, Lemma 21].

4. \[ \|AB - \bar{A}B\|_F = \|AB - \bar{A}B + \bar{A}\bar{B} - \bar{A}B\|_F \]
\[ \leq \|AB - \bar{A}B\|_F + \|\bar{A}\bar{B} - \bar{A}B\|_F \leq \|B\|\|A - \bar{A}\|_F + \|\bar{A}\|\|B - \bar{B}\|_F \]
\[ \leq (1 + \epsilon_B)\epsilon_A + \epsilon_B = \epsilon_A + \epsilon_B + \epsilon_A \epsilon_B. \]

\[ \square \]

1.4.3 Kronecker product and the vectorization of matrices

**Definition 7** (Kronecker Product). Let \(A \in \mathbb{R}^{\ell \times k}, B \in \mathbb{R}^{m \times n}\) be real matrices. The Kronecker product \(A \otimes B\) of \(A\) and \(B\) is given by

\[ A \otimes B := \begin{pmatrix} a_{11}B & \cdots & a_{1k}B \\ \vdots & \ddots & \vdots \\ a_{\ell 1}B & \cdots & a_{\ell k}B \end{pmatrix} \in \mathbb{R}^{\ell m \times kn}. \]

**Definition 8.** Let \(X \in \mathbb{R}^{m \times n}\) be a matrix. We define the vectorization of \(X\) as the \(mn \times 1\) vector obtained by stacking the columns of \(X\) on top of each other:

\[ \text{vec}(X) = (x_{11}, \ldots, x_{m1}, x_{12}, \ldots, x_{m2}, \ldots, x_{1n}, \ldots, x_{mn})^T \in \mathbb{R}^{mn}. \]

This transformation is a one-to-one mapping between the spaces \(\mathbb{R}^{m \times n}\) and \(\mathbb{R}^{mn}\). Most arithmetic operations on matrices behave equivalently for their vectorized counterpart. Let \(A, B\) be real matrices of the same dimensionality and \(\alpha \in \mathbb{R}\). Then

\[ \text{vec}(A + B) = \text{vec}(A) + \text{vec}(B), \]
\[ \text{vec}(\alpha A) = \alpha \text{vec}(A). \]

Additionally, the Frobenius inner product of two matrices is the Euclidean inner product of their vectorizations

\[ \langle \text{vec}(A), \text{vec}(B) \rangle = \text{Tr}(AB^T) = \langle A, B \rangle, \quad (1.4.6) \]
which implies that
\[ \| \text{vec}(A) \|_2 = \sqrt{\langle \text{vec}(A), \text{vec}(A) \rangle} = \sqrt{\langle A, A \rangle} = \| A \|_F. \tag{1.4.7} \]

These properties follow by a simple calculation. Lastly, we can express the vectorization of a matrix product as a linear transformation by using the Kronecker product.

**Lemma 6.** Let $A, B, C$ be real matrices. Then

1. $(A \otimes B)^T = A^T \otimes B^T$
2. $\text{rank}(A \otimes B) = \text{rank}(A) \text{rank}(B)$
3. If the product $ABC$ is defined then
\[ \text{vec}(ABC) = (C^T \otimes A) \text{vec}(B). \]

For a proof of the third point we refer the reader to [21, Theorem 10], which also includes a good overview over several other properties of the Kronecker in combination with the vec-operator and a proof for the remaining two points in Lemma 6.
Chapter 2

Problem specification

This chapter serves two purposes: First, it precisely describes the problem we try to solve together with our general conditions. Second, it introduces the notation that comes along with the problem and the involved quantities. We will consistently use the same notation throughout the upcoming chapters and therefore all upcoming results have to be seen in context of this chapter.

Definition 9. Let $f : \mathbb{R}^d \to \mathbb{R}$ be a function with a representation

$$f(x) = \sum_{\ell=1}^{m_1} h_\ell(b_\ell^T x) = \begin{pmatrix} g_1(a_1^T x) \\ g_2(a_2^T x) \\ \vdots \\ g_m(a_m^T x) \end{pmatrix},$$

such that the following properties hold:

1. $0 < m_1 \leq m \leq d$,
2. $a_1, \ldots, a_m \in \mathbb{R}^d$, $b_1, \ldots, b_{m_1} \in \mathbb{R}^{m_1}$ are two sets of independent vectors,
3. $\|a_i\|_2 \leq 1$ for all $i = 1, \ldots, m$ and $\|b_\ell\|_2 = 1$ for all $\ell = 1, \ldots, m_1$,
4. $g_1, \ldots, g_m, h_1, \ldots, h_{m_1} \in C^3(\mathbb{R})$ are univariate three times continuously differentiable functions,
5. $g'_i(0) \neq 0$ for all $i = 1, \ldots, m$, where $g'_i$ is the first derivative of $g_i$.
6. The vectors $a_1, \ldots, a_m$, $b_1, \ldots, b_{m_1}$ are quasi-orthogonal with $\epsilon_A = S(a_1, \ldots, a_m)$, $\epsilon_B = S(b_1, \ldots, b_\ell)$. Further denote by $b_{i\ell}$ the $i$th coordinate of the vector $b_\ell$, then

$$0 \leq \sum_{\ell=1}^{m_1} b_{i\ell}^2 < 1 - \left( 2\epsilon_A + 2\epsilon_B + (\epsilon_A + \epsilon_A \epsilon_B + \epsilon_A \epsilon_B) \frac{1}{1 - \epsilon_B} \right)$$

for all $i = 1, \ldots, m$, where

$$\epsilon_{AB} := \epsilon_A + \epsilon_B + \epsilon_A \epsilon_B.$$
We define $\mathcal{F}_d(m_1, m)$ as the class of functions $f : \mathbb{R}^d \to \mathbb{R}$ with a representation (2.0.1) such that the properties 1-6 hold for $m$ and $m_1$.

From this point on we introduce the following convention. If we work with a function $f \in \mathcal{F}_d(m_1, m)$, then the vectors $a_1, \ldots, a_m, b_1, \ldots, b_{m_1}$, constants $\epsilon_A, \epsilon_B, \epsilon_{AB}$ and functions $g_1, \ldots, g_m, h_1, \ldots, h_{m_1}$, refer to the quantities given in Definition [9]. This convention also includes all upcoming definitions w.r.t. these quantities.

For the rest of this section let $f$ be some function in $\mathcal{F}_d(m_1, m)$. Denote by $A \in \mathbb{R}^{d \times m}, B \in \mathbb{R}^{m \times m_1}$ the matrices

$$A := (a_1 \ldots | a_m), \quad B := (b_1 \ldots | b_{m_1}).$$

If not stated otherwise, $\bar{A}, \bar{B}$ will refer to the optimal orthogonal matrices of $A, B$ and $\bar{a}_i, \bar{b}_i$ will refer to the orthonormal columns of $\bar{A}, \bar{B}$. We introduce the following constants with respect to $g_1, \ldots, g_m, h_1, \ldots, h_{m_1}$:

$$\kappa_n := \max_{i=1, \ldots, m} \sup_{t \in [-1, 1]} |g_i^{(n)}(t)| \text{ for } n = 0, 1, 2, 3, \quad (2.0.3)$$

$$\eta_n := \sup_{\ell=1, \ldots, m_1} \sup_{t \in [-\kappa_0 \sqrt{m}, \kappa_0 \sqrt{m}]} |h_{\ell}^{(n)}(t)| \text{ for } n = 1, 2, 3, \quad (2.0.4)$$

where $g_i^{(n)}, h_{\ell}^{(n)}$ denotes the $n$-th derivative of $g_i, h_{\ell}$ for $n \geq 1$ and simply $g_i, h_{\ell}$ for $n = 0$.

We introduce $g : \mathbb{R}^m \to \mathbb{R}^m, h : \mathbb{R}^{m_1} \to \mathbb{R}^{m_1}$ given by

$$g(y) = g(\begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_m \end{pmatrix}) = \begin{pmatrix} g_1(y_1) \\ g_2(y_2) \\ \vdots \\ g_m(y_m) \end{pmatrix}, \quad (2.0.5)$$

$$h(z) = h(\begin{pmatrix} z_1 \\ z_2 \\ \vdots \\ z_{m_1} \end{pmatrix}) = \begin{pmatrix} h_1(z_1) \\ h_2(z_2) \\ \vdots \\ h_{m_1}(z_{m_1}) \end{pmatrix}. \quad (2.0.6)$$

With a slight abuse of notation we extend this definition to the derivatives, writing $g'$ for

$$g'(y) = g'(\begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_m \end{pmatrix}) = \begin{pmatrix} g_1'(y_1) \\ g_2'(y_2) \\ \vdots \\ g_m'(y_m) \end{pmatrix},$$

and we define $g'', g'''$, $h', h'', h'''$ respectively. Using this representation allows us to express the terms in (2.0.1) in a compact way:

$$\begin{pmatrix} g_1(a_1^T x) \\ g_2(a_2^T x) \\ \vdots \\ g_m(a_m^T x) \end{pmatrix} = g(A^T x).$$
For \( \mathbf{1} = [1, 1, \ldots, 1]^T \in \mathbb{R}^d \) we can now write \( f \) as

\[
    f(x) = \mathbf{1}^T h(B^T g(A^T x)).
\]

(2.0.7)

Our goal is to get a good approximation of \( A, B \) by only using evaluations of \( f \). To access \( a_1, \ldots, a_m \) and \( b_1, \ldots, b_m \), note that the gradient of \( f \) is given by

\[
    \nabla f(x) = \sum_{\ell=1}^{m_1} h'_\ell(b^T_\ell g(A^T x)) \sum_{i=1}^m b_{i\ell} g'_i(a^T_\ell x) a_i
\]

(2.0.8)

\[
    = \sum_{i=1}^m \left[ \sum_{\ell=1}^{m_1} h'_\ell(b^T_\ell g(A^T x)) b_{i\ell} \right] g'_i(a^T_\ell x) a_i.
\]

Thus \( \nabla f(x) \in \text{span} \{ a_1, \ldots, a_m \} \subset \mathbb{R}^d \). The Hessians \( H[f], (H[f])_{ij} = \frac{\partial^2 f}{\partial x_i \partial x_j} \) of \( f \) are given by

\[
    H[f](x) = \sum_{\ell=1}^{m_1} h''_\ell(b^T_\ell g(A^T x)) \sum_{i,j=1}^m b_{i\ell} b_{j\ell} g'_i(a^T_\ell x) g'_j(a^T_\ell x) a_i \otimes a_j
\]

(2.0.9)

\[
    + \sum_{\ell=1}^{m_1} h'_\ell(b^T_\ell g(A^T x)) \sum_{i=1}^m b_{i\ell} g''_i(a^T_\ell x) a_i \otimes a_i,
\]

where \( a_i \otimes a_j = a_i a_j^T \in \mathbb{R}^{d \times d} \) denotes the tensor product of the vectors \( a_i, a_j \). From (2.0.9) we can immediately deduce that

\[
    H[f](x) \in \text{span} \{ a_i \otimes a_j | i, j = 1, \ldots, m \},
\]

(2.0.10)

which is a subspace of the vectorspace of matrices in \( \mathbb{R}^{d \times d} \). Looking at the basis elements from (2.0.10) we demonstrate a useful identity applied to rank 1 matrices. Let \( i, i', j, j' \in 1, \ldots, m \). Then

\[
    \langle a_i \otimes a_j, a_{i'} \otimes a_{j'} \rangle = \text{Tr}(a_i \otimes a_j (a_{i'} \otimes a_{j'})^T)
\]

\[
    = \text{Tr}(a_i \otimes a_j a_{j'} \otimes a_{i'})
\]

\[
    = \langle a_i, a_{i'} \rangle \langle a_j, a_{j'} \rangle.
\]

The Hessians will always be symmetric. In fact, we can easily express \( H[f](x) \) as a sum of symmetric matrices:

\[
    H[f](x) = \sum_{\ell=1}^{m_1} h''_\ell(b^T_\ell g(A^T x)) \sum_{i \leq j} b_{i\ell} b_{j\ell} g'_i(a^T_\ell x) g'_j(a^T_\ell x) (a_i \otimes a_j + a_j \otimes a_i)
\]

(2.0.11)

\[
    + \sum_{\ell=1}^{m_1} h'_\ell(b^T_\ell g(A^T x)) \sum_{i=1}^m b_{i\ell} g''_i(a^T_\ell x) a_i \otimes a_i,
\]

and \( (a_i \otimes a_i)^T = (a_i \otimes a_i), (a_i \otimes a_j + a_j \otimes a_i)^T = (a_j \otimes a_i + a_i \otimes a_j) \). Thus

\[
    H[f](x) \in \text{span} \{ a_j \otimes a_i + a_i \otimes a_j | i, j = 1, \ldots, m; i \leq j \},
\]

(2.0.12)
which again leads to
\[
\dim \text{span}\{H[f](x) \mid x \in \mathbb{R}^d\} \leq \frac{m(m+1)}{2}.
\] (2.0.13)

By rewriting (2.0.9), we get
\[
H[f](x) = \sum_{\ell=1}^{m_1} h_\ell\left( b_\ell^T g(A^T x) \right) \sum_{i=1}^{m} b_{i\ell} g_i(a_i^T x) a_i \otimes a_i + \sum_{\ell=1}^{m_1} h_\ell\left( b_\ell^T g(A^T x) \right) \sum_{i=1}^{m} b_{i\ell} g_i(a_i^T x) a_i \otimes a_i.
\]

Let \( G_x \in \mathbb{R}^{m \times m} \) be the diagonal matrix (function) defined for every \( x \in \mathbb{R}^d \) as
\[
G_x := \begin{pmatrix}
g_1'(a_1^T x) \\
g_2'(a_2^T x) \\
\vdots \\
g_m'(a_m^T x)
\end{pmatrix}.
\]

Now, we can define
\[
v_{x,\ell} := AG_x b_{\ell} = \sum_{i=1}^{m} b_{i\ell} g_i(a_i^T x) a_i. \tag{2.0.14}
\]

Additionally, the second sum in our expression can be written as
\[
\sum_{\ell=1}^{m_1} h_\ell\left( b_\ell^T g(A^T x) \right) \sum_{i=1}^{m} b_{i\ell} g_i(a_i^T x) a_i \otimes a_i = \sum_{i=1}^{m} \sum_{\ell=1}^{m_1} h_\ell\left( b_\ell^T g(A^T x) \right) b_{i\ell} g_i(a_i^T x) a_i \otimes a_i = \sum_{i=1}^{m} \left[ Bh_i'(B^T g(A^T x)) \right] g_i(a_i^T x) a_i \otimes a_i.
\]

Finally, we introduce the diagonal matrices \( S_x \in \mathbb{R}^{m_1 \times m_1}, T_x \in \mathbb{R}^{m \times m} \) by giving their diagonal elements:
\[
s_{x,\ell\ell} := h_\ell''(b_\ell^T g(A^T x)) \quad \text{for } \ell = 1, \ldots, m_1, \tag{2.0.15}
\]
\[
t_{x,ii} := \left[ Bh_i'(B^T g(A^T x)) \right] g_i''(a_i^T x) \quad \text{for } i = 1, \ldots, m. \tag{2.0.16}
\]

Thus, the Hessian can be read as
\[
H[f](x) = \sum_{\ell=1}^{m_1} s_{x,\ell\ell} v_{x,\ell} \otimes v_{x,\ell} + \sum_{i=1}^{m} t_{x,ii} a_i \otimes a_i. \tag{2.0.17}
\]
Further calculation yields
\[
\sum_{\ell=1}^{m_1} s_{x,\ell} v_{x,\ell} \otimes v_{x,\ell} = AG_x \left( \sum_{\ell=1}^{m_1} s_{x,\ell} b_{\ell} \otimes b_{\ell} \right) G_x A^T \\
= AG_x B S_x B^T G_x A^T.
\]
Applying the same argument to the sum of \( a_i \otimes a_i' \)'s and denoting \( V_x := AG_x B \), we get
\[
H[f](x) = V_x S_x V_x^T + A T_x A^T. \tag{2.0.18}
\]
Since \( H[f](x) = A \left( G_x B S_x B^T G_x + T_x \right) A^T \), we have \( \text{rank} (H[f](x)) \leq m \) with equality if and only if \( G_x B S_x B^T G_x + T_x \) has full rank \( m \).

By (2.0.17) there is a decomposition of \( H[f](x) \) that consists of \( m + m_1 \) rank-one matrices. This does not imply that \( \text{span} \{ H[f](x) \mid x \in \mathbb{R}^d \} \) is a \( m + m_1 \) dimensional subspace, since the elements \( v_{x,1}^{\otimes 2}, \ldots, v_{x,m_1}^{\otimes 2} \) depend on \( x \). Nevertheless, our intuition tells us that the Hessians might concentrate around a space that is related to (2.0.17) and has dimension \( m + m_1 \). The following space will be our primary interest throughout the next chapter:
\[
L := \text{span} \{ a_i^{\otimes 2}, v_i^{\otimes 2} \mid i = 1, \ldots, m, l = 1, \ldots, m_1 \}, \tag{2.0.19}
\]
where \( v_i := v_{0,l} = AG_0 b_i \). Since \( f \in \mathcal{F}_d(m_1, m) \) we have \( g'(0) \neq 0 \), which makes \( G_0 \) non-singular. First let us show some properties of the elements in \( L \).

**Lemma 7.** The matrices in \( \{ a_i^{\otimes 2} \mid i = 1, \ldots, m \} \) are linear independent.

It is well known that the symmetric tensor product of independent vectors yields independent tensors. We give a short proof that demonstrates some of the previously introduced tools.

**Proof.** This follows directly from the properties of the Kronecker product defined in section 1.4.3. Assume there exists a \( \lambda \in \mathbb{R}^m \) such that
\[
\sum_{i=1}^m \lambda_i a_i \otimes a_i = 0.
\]
Denote by \( D_\lambda \) the diagonal matrix with \( \lambda \) as it’s diagonal. Then
\[
\text{vec} \left( \sum_{i=1}^m \lambda_i a_i \otimes a_i \right) = \text{vec} (A D_\lambda A^T) \\
= A \otimes A \text{vec} (D_\lambda).
\]
By Lemma 6 we get \( \text{rank} A \otimes A = (\text{rank} A)^2 = m^2 \). If the linear combination was 0, it’s vectorization has to be 0. The matrix \( A \otimes A \) has full rank, hence
\[
A \otimes A \text{vec}(D_\lambda) = 0 \iff D_\lambda = 0.
\]
This implies that \( \lambda \) was 0, therefore the matrices are independent. \( \square \)
By the exact same argument we can show that the matrices in \( \{ v_{\ell}^{\otimes 2} \mid \ell = 1, \ldots, m_1 \} \) are linear independent, since \( AG_0B \) has full column rank. One can easily see that this does not extend to \( \{ a_i^{\otimes 2}, v_{\ell}^{\otimes 2} \mid i = 1, \ldots, m; \ell = 1, \ldots, m_1 \} \) in general. In fact, the vectors \( v_{\ell} \) are linear combinations of the vectors \( a_1, \ldots, a_m \). If the properties in Definition 9 would not hold (in particular property 6), we could pick \( b_1 = [1, 0, \ldots, 0]^T \) and get

\[
v_1 \otimes v_1 = (AG_0b_1) \otimes (AG_0b_1)
= (Ag_1'(0)b_1) \otimes (Ag_1'(0)b_1)
= g_1'(0)^2 a_1 \otimes a_1.
\]

However, the next result shows that the properties in Definition 9 are sufficient for the linear independence.

**Theorem 1.** In the setting above, the properties 1-6 from Definition 9 guarantee that the matrices in

\[
\{ a_i^{\otimes 2}, v_{\ell}^{\otimes 2} \mid i = 1, \ldots, m; \ell = 1, \ldots, m_1 \}
\tag{2.0.20}
\]

are linear independent. Therefore \( L \) has dimension \( m + m_1 \) and \( (2.0.20) \) is a minimal basis of \( L \).

**Proof.** Denote by \( \bar{A}, \bar{B} \) the optimal orthogonal matrices w.r.t. \( A, B \). Assume that the elements are not independent, then there are vectors \( \lambda \in \mathbb{R}^m, \mu \in \mathbb{R}^{m_1} \) with

\[
\sum_{i=1}^{m} \lambda_i a_i^{\otimes 2} + \sum_{\ell=1}^{m_1} \mu_\ell v_{\ell}^{\otimes 2} = 0.
\]

We already established that the \( a_i^{\otimes 2} (, v_{\ell}^{\otimes 2}) \) are independent. So the independence can only be destroyed by mixing the \( a_i^{\otimes 2} \) with \( v_{\ell}^{\otimes 2} \), which leads to \( \lambda, \mu \neq 0 \). Let \( D_\lambda, D_\mu \) be the diagonal matrices with \( \lambda, \mu \) as diagonal. Now we can use the representation from before to express the linear combination as

\[
\sum_{i=1}^{m} \lambda_i a_i^{\otimes 2} + \sum_{\ell=1}^{m_1} \mu_\ell v_{\ell}^{\otimes 2} = AD_\lambda A^T + AG_0BD_\mu B^T G_0 A^T.
\tag{2.0.21}
\]

The first observation is that

\[
\text{vec}(AD_\lambda A^T + AG_0BD_\mu B^T G_0 A^T) = 0
\]
\[
\iff (A \otimes A) \text{vec}(D_\lambda + G_0BD_\mu B^T G_0) = 0.
\]

Due to our assumptions on \( A \), the matrix \( (A \otimes A) \) has full column rank. Hence, the last equality implies

\[
\text{vec}(D_\lambda + G_0BD_\mu B^T G_0) = 0
\]
\[
\iff G_0BD_\mu B^T G_0 = -D_\lambda
\]
\[
\iff BD_\mu B^T = -G_0^{-1} D_\lambda G_0^{-1}.
\]

18
Now consider the following inner product
\[ G_0 BD \mu B^T G_0 = G_0^2 BD \mu B^T. \] (2.0.22)

W.l.o.g we can assume \( \|D_\lambda\|_F = 1 \), otherwise we could just multiply the linear combination by \( \frac{1}{\|D_\lambda\|_F} \). This together with (2.0.22) yields \( \|G_0^2 BD \mu B^T\|_F = \|D_\lambda\|_F = 1 \). By the quasi-orthogonality of \( B \) we can further estimate
\[ \frac{1}{1 + \epsilon_B} \leq \|G_0^2 BD \mu\|_F \leq \frac{1}{1 - \epsilon_B}. \]

Since the linear combination vanishes, the inner product of \( AD_\lambda A^T + AG_0 BD \mu B^T G_0 A^T \) with any matrix in \( \mathbb{R}^{d \times d} \) has to be zero as well. Hence
\[ \langle AD_\lambda A^T + AG_0 BD \mu B^T G_0 A^T, \bar{A}D_\lambda(I_m - \bar{B}\bar{B}^T)\bar{A}^T \rangle = 0. \] (2.0.23)

Now consider the following inner product
\[
\langle \bar{A}D_\lambda \bar{A}^T + \bar{A}G_0^2 BD \mu \bar{B}^T \bar{A}^T - AD_\lambda A^T - AG_0 BD \mu B^T G_0 A^T, \bar{A}D_\lambda(I_m - \bar{B}\bar{B}^T)\bar{A}^T \rangle
\]

\[
= \langle \bar{A}D_\lambda \bar{A}^T + \bar{A}G_0^2 BD \mu \bar{B}^T \bar{A}^T, \bar{A}D_\lambda(I_m - \bar{B}\bar{B}^T)\bar{A}^T \rangle - 0
\]

\[
= \text{Tr} \left( \langle \bar{A}D_\lambda \bar{A}^T + \bar{A}G_0^2 BD \mu \bar{B}^T \bar{A}^T \rangle \bar{A}(I_m - \bar{B}\bar{B}^T)D_\lambda \bar{A}^T \right)
\]

\[
= \text{Tr} \left( \langle D_\lambda + G_0^2 BD \mu \bar{B}^T \rangle (I_m - \bar{B}\bar{B}^T)D_\lambda \right)
\]

\[
= \text{Tr} \left( D_\lambda D_\lambda - D_\lambda \bar{B}\bar{B}^T D_\lambda + G_0^2 BD \mu \bar{B}^T D_\lambda - G_0^2 BD \mu \bar{B}^T D_\lambda \right)
\]

\[
= \|D_\lambda(I_m - \bar{B}\bar{B}^T)\|_F^2.
\]

Using (2.0.22) and applying the Cauchy-Schwarz inequality to the inner product yields the following upper bound:
\[
\left| \langle \bar{A}D_\lambda \bar{A}^T + \bar{A}G_0^2 BD \mu \bar{B}^T \bar{A}^T - AD_\lambda A^T - AG_0^2 BD \mu B^T A^T, \bar{A}D_\lambda(I_m - \bar{B}\bar{B}^T)\bar{A}^T \rangle \right|
\]

\[
\leq \| \bar{A}D_\lambda \bar{A}^T + \bar{A}G_0^2 BD \mu \bar{B}^T \bar{A}^T - AD_\lambda A^T - AG_0^2 BD \mu B^T A^T \|_F \| \bar{A}D_\lambda(I_m - \bar{B}\bar{B}^T)\bar{A}^T \|_F
\]

\[
\leq \left( \| \bar{A}D_\lambda \bar{A}^T - AD_\lambda A^T \|_F + \| \bar{A}G_0^2 BD \mu \bar{B}^T \bar{A}^T - AG_0^2 BD \mu B^T A^T \|_F \right) \| D_\lambda(I_m - \bar{B}\bar{B}^T) \|_F
\]

\[
\leq \left( 2\epsilon_A + \epsilon_A + \epsilon_A \epsilon_B \right) \| G_0^2 BD \mu \|_F \| D_\lambda(I_m - \bar{B}\bar{B}^T) \|_F
\]

\[
\leq \left( 2\epsilon_A + \epsilon_A + \epsilon_A \epsilon_B \right) \frac{1}{1 - \epsilon_B} \| D_\lambda(I_m - \bar{B}\bar{B}^T) \|_F.
\]
Hence, by (2.0.23), the scalar product

\[
\langle AD\lambda \tilde{A}^T + \tilde{A}BD\mu \tilde{B}^T \tilde{A}^T - AD\lambda \tilde{A}^T - ABD\mu \tilde{B}^T \tilde{A}^T, \tilde{A}D\lambda (I_m - \tilde{B}\tilde{B}^T)\tilde{A}^T \rangle = \|D\lambda (I_m - \tilde{B}\tilde{B}^T)\|_F^2
\]

is bounded from above by \(2\varepsilon_A + (\varepsilon_A + \varepsilon_A \varepsilon_{AB} + \varepsilon_{AB}) \frac{1}{1 - \varepsilon_B} \|D\lambda (I_m - \tilde{B}\tilde{B}^T)\|_F\). This would imply that

\[
\|D\lambda (I_m - \tilde{B}\tilde{B}^T)\|_F \leq \left(2\varepsilon_A + (\varepsilon_A + \varepsilon_A \varepsilon_{AB} + \varepsilon_{AB}) \frac{1}{1 - \varepsilon_B}\right).
\]

(2.0.24)

Due to \(\|D\lambda\|_F = \|\lambda\|_2 = 1\) we can further estimate

\[
\|D\lambda (I_m - \tilde{B}\tilde{B}^T)\|_F = \left(\text{Tr}(D\lambda^2(I_m - \tilde{B}\tilde{B}^T))\right)^{\frac{1}{2}} \geq \left(\sum_{i=1}^m \lambda_i^2 (1 - \sum_{\ell=1}^{m_1} \tilde{b}_{i\ell}^2)\right)^{\frac{1}{2}} \geq \min_i \left|1 - \sum_{\ell=1}^{m_1} \tilde{b}_{i\ell}^2 \bigg| - \sum_{\ell=1}^{m_1} \tilde{b}_{i\ell}^2 - \tilde{b}_{i\ell}^2 \right|.
\]

Assume the last expression is minimal for \(i = i_0\):

\[
\min_i \left|1 - \sum_{\ell=1}^{m_1} \tilde{b}_{i\ell}^2 \bigg| - \sum_{\ell=1}^{m_1} \tilde{b}_{i\ell}^2 - \tilde{b}_{i\ell}^2 \right| \geq \left|1 - \sum_{\ell=1}^{m_1} \tilde{b}_{i\ell}^2 \right|.
\]

The right part of the last equality \(\sum_{\ell=1}^{m_1} \tilde{b}_{i\ell}^2 - \tilde{b}_{i\ell}^2\) is the difference of the squared elements in the \(i_0\)-th row of \(B\) and \(\tilde{B}\). To not confuse the row with the column vector, we denote the \(i_0\)-th rows of \(B\), \(\tilde{B}\) as \(b_{i_0}, \tilde{b}_{i_0}\), such that

\[
\sum_{\ell=1}^{m_1} \tilde{b}_{i_0\ell}^2 - \tilde{b}_{i_0\ell}^2 = \sum_{\ell=1}^{m_1} b_{i_0\ell}^2 - \sum_{\ell=1}^{m_1} \tilde{b}_{i_0\ell}^2 = \|b_{i_0}\|_2^2 - \|\tilde{b}_{i_0}\|_2^2
\]

\[
= \left(\|b_{i_0}\|_2^2 - \|\tilde{b}_{i_0}\|_2^2\right) \left(\|b_{i_0}\|_2^2 + \|\tilde{b}_{i_0}\|_2^2\right) \leq 2\varepsilon_B.
\]

Together with the inequality on \(1 - \sum_{\ell=1}^{m_1} b_{i_0\ell}^2\) from assumption (2.0.2) we get

\[
\|D\lambda (I_m - \tilde{B}\tilde{B}^T)\|_F \geq \left|1 - \sum_{\ell=1}^{m_1} b_{i_0\ell}^2 \right| - \left|1 - \sum_{\ell=1}^{m_1} \tilde{b}_{i_0\ell}^2 - \tilde{b}_{i_0\ell}^2 \right|
\]

\[
> \left(2\varepsilon_A + (\varepsilon_A + \varepsilon_A \varepsilon_{AB} + \varepsilon_{AB}) \frac{1}{1 - \varepsilon_B}\right) + 2\varepsilon_B - \left|\sum_{\ell=1}^{m_1} \tilde{b}_{i_0\ell}^2 - \tilde{b}_{i_0\ell}^2 \right|
\]

\[
\geq \left(2\varepsilon_A + (\varepsilon_A + \varepsilon_A \varepsilon_{AB} + \varepsilon_{AB}) \frac{1}{1 - \varepsilon_B}\right).
\]

This is a contradiction to (2.0.23). Therefore, we can conclude that there is no \(\lambda, \mu \neq 0\) such that

\[
\sum_{i=1}^m \lambda_i a_i^2 + \sum_{\ell=1}^{m_1} \mu_{i\ell} v_i^2 = 0.
\]

\[\square\]
Overview over the upcoming chapters

By Definition 9 and the construction above we can find the subspace $L \subset \mathbb{R}^{d \times d}$ defined in 2.0.19 for every $f \in \mathcal{F}_d(m_1, m)$. Since these functions fulfill the property specified in 2.0.2, the previous theorem implies that the matrices in

$$\{a_i \otimes^2, v_i \otimes^2 \mid i = 1, \ldots, m; l = 1, \ldots, m_1\} \tag{2.0.25}$$

are independent and form a minimal basis of $L$.

The remaining part of this thesis will be focused on the recovery of the matrices in (2.0.25) for an arbitrary $f \in \mathcal{F}_d(m_1, m)$. In chapter 3 we present an algorithm, which approximates $L$ by a space of the same dimension, and develop several bounds on the approximation error. The first version of our algorithm works under the assumption that the second derivative of $f$ is known. This will be complemented by the results of chapter 4, where we replace the Hessian by a finite-difference approximation. Incorporating finite-differences yields an algorithm that only relies on evaluations of $f$ itself. Finally, in chapter 5 we address a method that recovers the vectors $a_1, \ldots, a_m, v_1, \ldots, v_m$ from the approximation of $L$. The last chapter is given by several numerical experiments where we apply all these methods. The identification of $a_i, v_i$ will only be possible up to a sign.

Additionally, we will not discuss the identification of the functions $g_1, \ldots, g_m, h_1, \ldots, h_m$. Therefore, we are missing a crucial part in the reconstruction of the vectors $b_1, \ldots, b_m$. Inferring $b_\ell$ from $v_\ell$ can only be done if $G_0$ is known:

$$v_\ell = AG_0 b_\ell.$$ 

Hence, to be able to recover $b_1, \ldots, b_m$, we have to assume that $g_1'(0), \ldots, g_m'(0)$ is known.
Chapter 3

Concentration of the Hessians around $L$

Throughout this chapter we assume that $f \in \mathcal{F}_d(m_1, m)$ according to Definition 9 and that $L$ is the matrix space defined in (2.0.19). We introduce an algorithm that approximates $L$ from random evaluations of the Hessian of $f$. The final goal is a method to recover $A, B$ by only using evaluations of $f$, which means that $H[f](x)$ is usually not known. In chapter 4 we will show how we can approximate $H[f](x)$ by second order finite-differences and how it affects our approximation of $L$. For now we assume that the second derivative $H[f](x)$ of $f$ is available.

3.1 Approximating $L$ by $L_{PCA}$

Let’s assume we have $m_x \in \mathbb{N}$ different points $x_1, \ldots, x_{m_x} \in \mathbb{R}^d$ and $H[f](x_1), \ldots, H[f](x_{m_x}) \in \mathbb{R}^{d \times d}$ are known. Now we can construct the matrix $M$ given by

$$
M := \text{vec}(H[f](x_1)) | \ldots | \text{vec}(H[f](x_{m_x})) \in \mathbb{R}^{d^2 \times m_x}.
$$

(3.1.1)

If $m_x$ is large and the $x_i$ are well distributed, the column space $\text{range}(M)$ of $M$ will eventually come very close the space spanned by the vectorized Hessians. To make sure the points $x_1, \ldots, x_{m_x}$ are well distributed, we will draw from the uniform distribution on the unit sphere

$$
\mathbb{S}^{d-1} := \{x \in \mathbb{R}^d \mid \|x\|_2 = 1\}.
$$

(3.1.2)

We get the approximation $L_{PCA} \approx L$ by setting $L_{PCA}$ to the column space of the first $m + m_1$ left singular vectors of $M$. To be precise, let

$$
M = \begin{pmatrix}
U_1 & U_2
\end{pmatrix}
\begin{pmatrix}
\Sigma_1 & 0 \\
0 & \Sigma_2
\end{pmatrix}
\begin{pmatrix}
V_1^T \\
V_2^T
\end{pmatrix}
$$

(3.1.3)

be the singular value decomposition of $M$, where $U_1 = (u_1 | \ldots | u_{m+m_1}) \in \mathbb{R}^{d^2 \times (m+m_1)}$. Then we set $L_{PCA}$ to the column space $\text{range}(U_1)$ of $U_1$. This procedure is summarized in Algorithm 1.
Remark. The space \( L \subset \mathbb{R}^{d \times d} \) is a subspace of matrices, while \( L_{PCA} \subset \mathbb{R}^{d^2} \) is a vector subspace. We are aware that this might cause confusion and will address our representation of matrix subspaces at the beginning of Section 3.3.

Algorithm 1: Dimension reduction on the span of the Hessians

Input: \( f \in F_d(m_1, m) \)

Parameters: \( m_x \)

Output: \( L_{PCA} \)

1 begin
2 Sample \( x_1, \ldots, x_{m_x} \) uniformly from \( S^{d-1} \);
3 \( M \leftarrow (\text{vec}(H[f](x_1)), \ldots, \text{vec}(H[f](x_{m_x}))); \)
4 \( (U_1 \ U_2) \Sigma V^T \leftarrow \text{SVD}(M); \)
5 \( L_{PCA} \leftarrow \text{range}(u_1 | \ldots | u_{m+m_1}); \)
6 end

The algorithm has to be seen as a general framework. It can be separated into two steps. First, we evaluate the Hessians such that \( \text{range}(M) \) is a rich space, hopefully containing enough information to reconstruct the directions \( a_1, \ldots, a_m, v_1, \ldots, v_L \). This clearly depends on the choice of our distribution. Drawing \( x_1, \ldots, x_{m_x} \) uniformly from the sphere has several benefits which will be discussed in the next section. Nevertheless, it is important to address that we do not claim that this distribution is optimal. Finding a better way to sample the Hessians depending on the structure of the problem is not a focus of this thesis, but we will demonstrate some ideas regarding this topic in the numerical part. One simple example is to change the radius of the sphere, which can have a large impact on the final result (see chapter 6).

The second step consists of a dimensionality reduction based on the singular value decomposition of \( M \). Picking the first \( m + m_1 \) singular vectors is a natural decision, we simply choose to approximate the \( m + m_1 \) dimensional space \( L \) by a \( m + m_1 \) dimensional space. Deviating from this choice would automatically lead to a difference between \( L \) and \( L_{PCA} \).

3.2 Uniform distribution on the unit-sphere

Before we study the result of Algorithm 1, we want to highlight some properties of the uniform spherical distribution. The first is the rotational invariance. To be more precise, let \( X \sim \text{Unif}(S^{d-1}) \) be a random vector distributed uniformly on \( S^{d-1} \). Define

\[
\mu_{S^{d-1}}(E) := \mathbb{P}(X \in E)
\]

for any Borel measurable set \( E \subset \mathbb{R}^d \). This yields a probability measure \( \mu_{S^{d-1}} \) on \( \mathbb{R}^d \). Since the sphere is invariant under rotations, we have

\[
O \cdot X \sim \text{Unif}(S^{d-1})
\]

23
for any unitary matrix $O \in \mathbb{R}^{d \times d}$. A common expression that occurs when we sample $f$ or its derivatives from the spherical distribution is the linear transformation $A^T X$ with a matrix in $\mathbb{R}^{d \times m}$. By our assumption $A$ is quasi-orthogonal, i.e. there is an orthogonal matrix $\bar{A}$ such that $A \approx \bar{A}$. Additionally,
\[ \|A^T X - \bar{A}^T X\|_2 \leq \|A - \bar{A}\| \|X\|_2 \leq \epsilon_A, \]
which allows us to study the distribution of $\bar{A}^T X$ as a proxy for $A^T X$, if we accept a small error depending on $\epsilon_A$. The rotational invariance together with the orthogonality of $\bar{A}$ yields
\[ \bar{A}^T X \overset{d}{=} \bar{A}^T O X. \quad (3.2.1) \]
Here $\overset{d}{=} \text{ stands for equivalence in distribution. By setting } O = (\bar{A} | \bar{A}_\perp), \text{ with } \bar{A}_\perp \text{ having orthonormal columns to all the } \bar{a}_i, \text{ (3.2.1) reduces to } \]
\[ \bar{A}^T X \overset{d}{=} [X_1, ..., X_m]^T. \]
Hence $\bar{A}^T X$ has the same distribution as the first $m$ marginals of $X$. A well known method to draw a vector $x \in \mathbb{R}^d$ uniformly from the unit sphere is to construct it by drawing $d$ standard normal variables.

**Lemma 8.** Let $Z_1, ..., Z_d \sim \mathcal{N}(0, 1)$ be independent standard normal variables and $Z := [Z_1, ..., Z_d]^T$. Then
\[ \frac{Z}{\|Z\|_2} \sim \text{Unif}(S^{d-1}). \quad (3.2.2) \]

**Lemma 9.** Suppose $X \sim \text{Unif}(S^{d-1})$, $\bar{A} \in \mathbb{R}^{d \times m}$ orthogonal, and $d > m > 0$. Then
\[ \|\bar{A}^T X\|_2^2 \sim \text{Beta} \left( \frac{m}{2}, \frac{d - m}{2} \right). \]

*For a definition of the Beta-distribution see Appendix (A.1)*

**Proof.** We already established that we can assume w.l.o.g $\bar{A}^T X = [X_1, ..., X_m]^T =: \hat{X}$. From Lemma 8 we know the distribution of the elements of $X$ is given by
\[ X_i \sim \frac{Z_i}{\sqrt{Z_1^2 + \cdots + Z_d^2}} \]
with independent $Z_1, ..., Z_d \sim \mathcal{N}(0, 1)$. Thus
\[ \|\hat{X}\|_2^2 = X_1^2 + \cdots + X_m^2 \sim \frac{Z_1^2 + \cdots + Z_m^2}{Z_1^2 + \cdots + Z_d^2}. \]
Set \( Q_1 = Z_1^2 + \ldots + Z_m^2, Q_2 = Z_{d-m+1}^2 + \ldots + Z_d^2 \), then \( Q_1, Q_2 \) are independent and
\[
Q_1 \sim \chi^2(m), \quad Q_2 \sim \chi^2(d - m),
\]
where \( \chi^2(k) \) denotes the chi-squared distribution with \( k \) degrees of freedom. We can then see that the chi-squared distribution is equal to the gamma distribution \([2, page 71]\) and by an argument in \([2, page 59]\) follows
\[
\| \bar{A}^T X \|_2^2 = \frac{Q_1}{Q_1 + Q_2} \sim \text{Beta} \left( \frac{m}{2}, \frac{d - m}{2} \right). \tag{3.2.4}
\]

In the next part we will introduce sub-gaussian random variables, which are characterized by a fast tail decay. There are several equivalent definitions of sub-gaussian random variables (see for example \([17, Section 5.2.3]\)), one is given as follows.

**Definition 10 (Sub-gaussian).** Let \( Y \) be a random variable with mean \( \mu \), we say \( Y \) is sub-gaussian if there exists a \( \sigma > 0 \) such that
\[
\mathbb{E} \left[ \exp (t(Y - \mu)) \right] \leq \exp \left( \frac{t^2 \sigma^2}{2} \right) \tag{3.2.5}
\]
for all \( t \in \mathbb{R} \).

**Definition 11 (Sub-gaussian norm).** Let \( Y \) be a sub-gaussian random variable. The sub-gaussian norm of \( Y \) is defined as
\[
\| Y \|_{\psi_2} := \inf \{ t > 0 \mid \mathbb{E} \exp (Y^2/t^2) \leq 2 \}.
\]

**Lemma 10 (cf. \([17, p. 11]\)).** If \( Y \) is sub-gaussian and (3.2.5) holds for \( \sigma^2 > 0 \), then
\[
\| Y - \mathbb{E}[Y] \|_{\psi_2}^2 \leq C \sigma^2,
\]
for some constant \( C > 0 \).

The beta distribution is bounded, therefore it is also sub-gaussian. We are interested in the minimal \( \sigma \) such that (3.2.5) is true for a beta-distributed random variable. This is explored in the work \([10]\), one of the results gives a upper bound for the optimal \( \sigma^2 \).

**Theorem 2 (cf. \([10, Theorem 1]\)).** Let \( Y \sim \text{Beta}(\alpha, \beta) \) with \( \alpha, \beta > 0 \) and \( \sigma^2_{\text{opt}}(\alpha, \beta) \) be the minimal \( \sigma^2 > 0 \) such that (3.2.3) holds. Then
\[
\sigma^2_{\text{opt}}(\alpha, \beta) \leq \frac{1}{4(\alpha + \beta + 1)}. \tag{3.2.6}
\]

Applying these results to \( \| \bar{A}^T X \|_2^2 \) yields
Corollary 1. Let $Y := \|\bar{A}^T X\|_2^2$, where $X \sim \text{Unif}(S^{d-1})$, $\bar{A} \in \mathbb{R}^{d \times m}$ orthogonal. Then

$$\|Y - \mathbb{E}[Y]\|_{\psi_2}^2 \leq \frac{C}{2d + 4},$$  \hfill (3.2.7)

for some constant $C > 0$.

Proof. Applying Lemma 9 yields $Y = \|\bar{A}^T X\|_2^2 \sim \text{Beta}(m/2, (d - m)/2)$. By Theorem 2 follows

$$\mathbb{E}[\exp(t(Y - \mathbb{E}[Y]))] \leq \exp\left(\frac{t^2 \sigma^2}{2}\right)$$

for

$$\sigma^2 \leq \frac{1}{4 \left(\frac{m}{2} + \frac{d-m}{2} + 1\right)} = \frac{1}{2d + 4}. \hfill (3.2.8)$$

Hence, by Lemma 10 follows

$$\|Y - \mathbb{E}[Y]\|_{\psi_2}^2 \leq C\sigma^2 = \frac{C}{2d + 4},$$

for some constant $C > 0$.

Lemma 11. Let $X \sim \text{Unif}(S^{d-1})$ and $\bar{A} \in \mathbb{R}^{d \times m}$ be an orthogonal matrix. Then

$$\mathbb{E}[\|\bar{A}^T X\|_\infty^2] \leq \left(\frac{2 \log(m)}{d}\right)^\frac{1}{2} + \frac{1}{d}. \hfill (3.2.9)$$

Proof. Let $X \sim \text{Unif}(S^{d-1})$ and $\bar{A} \in \mathbb{R}^{d \times m}$ be orthogonal. As noted before, $\bar{A}^T X$ follows the same distribution as the first $m$ coordinates of $X$. Set $Z := \|[X_1, \ldots, X_m]^T\|_\infty^2$. Then

$$Z = \max_{i \leq m} |X_i|^2.$$ 

Let $t > 0$, then $x \mapsto \exp(tx)$ is a convex and increasing function. Therefore, we can use Jensen’s inequality

$$\exp(t\mathbb{E}[Z]) \leq \mathbb{E}[\exp(tZ)] \leq \max_{i \leq m} \mathbb{E}[\exp(tX_i^2)]$$

$$\leq \sum_{i=1}^m \mathbb{E}[\exp(tX_i^2)] \leq m \mathbb{E}[\exp(tX_1^2)]$$

$$= m \mathbb{E}\left[\exp\left(tX_1^2 - \frac{t}{d}\right)\right] \exp\left(\frac{t}{d}\right).$$

By the theory above we know that $X_1^2 \sim \text{Beta}(1/2, (d - 1)/2)$ with expectation $1/d$. Furthermore, we can apply Theorem 2 to estimate

$$m \mathbb{E}\left[\exp\left(tX_1^2 - \frac{t}{d}\right)\right] \exp\left(\frac{t}{d}\right) \leq m \exp\left(\frac{t^2}{4d} + \frac{t}{d}\right).$$

26
Solving this expression for $E[Z]$ yields

$$E[Z] \leq \frac{\log(me^{\frac{t^2}{4t}})}{t} = \frac{\log m}{t} + \left(\frac{t/4 + 1}{d}\right).$$

Taking the derivative with respect to $t$ to minimize the upper bound:

$$0 = -\log(m)t^{-2} + \left(\frac{1}{4d}\right)$$

$\Leftrightarrow t = \pm 2\sqrt{\log(m)d}.$

Since $t > 0$, we conclude $t = 2\sqrt{\log(m)d}$, which leads to

$$E[Z] \leq \sqrt{\frac{2\log m}{d}} + \frac{1}{d}.$$

\[\square\]

### 3.3 Concentration of $L_{PCA}$ around $L$

As mentioned before, the space $L$ is a space of matrices, while $L_{PCA}$ is a subspace of $\mathbb{R}^{d^2}$. However, we can compare those spaces by either vectorizing the elements in $L$

$$L^{vec} := \text{span}\{\text{vec}(a_1^{\otimes 2}), \ldots, \text{vec}(a_m^{\otimes 2}), \text{vec}(v_1^{\otimes 2}), \ldots, \text{vec}(v_m^{\otimes 2})\} \subset \mathbb{R}^{d^2} \quad (3.3.1)$$

or by mapping $L_{PCA} = \text{span}(u_1, \ldots, u_{m+m_1})$ to the space

$$L_{PCA}^{mat} = \text{span}\{\text{mat}_{d\times d}(u_1), \ldots, \text{mat}_{d\times d}(u_{m+m_1})\} \subset \mathbb{R}^{d\times d}, \quad (3.3.2)$$

where $\text{mat}_{d\times d} : \mathbb{R}^{d^2} \rightarrow \mathbb{R}^{d\times d}$ is defined as the inverse of the vec-operator from Definition [8]. When we say that $L$ and $L_{PCA}$ are close, we mean that $L^{vec}$ and $L_{PCA}$ (or equivalently $L$ and $L_{PCA}^{mat}$) are close.

The distance between two spaces will be measured as the distance of their orthogonal projections. For more details on orthogonal projections we refer to [6, 4.4.4]. Any orthogonal projection onto a linear subspace of $\mathbb{R}^{n}$ can be uniquely represented by a matrix. In this case we will identify an orthogonal projection by its projection matrix. As an example we can use $L_{PCA}$, which is a subspace of $\mathbb{R}^{d^2}$. The orthogonal projection onto $L_{PCA}$ is given by a matrix $P_{L_{PCA}} \subset \mathbb{R}^{d^2\times d^2}$. In this case we even know that

$$P_{L_{PCA}} = U_1U_1^T, \quad (3.3.3)$$

where $U_1$ was defined in [3.1.3]. This is due to the fact that, by our construction, the columns of $U_1$ form an orthonormal basis of $L_{PCA}$. This representation also helps us to
compare subspaces. We can simply measure the distance of their orthogonal projections under some matrix norm.

Let $P_L$ be the orthogonal projection onto $L$, and suppose $x_1, \ldots, x_{m_x}$ are drawn uniformly from the unit sphere $\mathbb{S}^{d-1}$ at random. According to Algorithm 1, we construct $L_{PCA}$ as the span of the first $m + m_1$ left singular vectors of the matrix

$$M = [\text{vec}(H[f](x_1)) | \ldots | \text{vec}(H[f](x_{m_x}))] \in \mathbb{R}^{d^2 \times m_x}. \quad (3.3.4)$$

We further denote

$$\hat{M} = [\text{vec}(P_L H[f](x_1)) | \ldots | \text{vec}(P_L H[f](x_{m_x}))] \in \mathbb{R}^{d^2 \times m_x} \quad (3.3.5)$$

and set $\hat{L} = \text{range}(\hat{M})$. Note that $\hat{L} \subseteq L^{vec}$ (cf. 3.3.1), therefore

$$\dim(\hat{L}) \leq \dim(L^{vec}) \leq m + m_1.$$ 

We want to show that $L_{PCA}$ concentrates around $L^{vec}$. This will be done in two steps. First, we develop upper bounds for the distance of $L_{PCA}$ and $\hat{L}$. Second, we show that $\hat{L} = L^{vec}$ if $L_{PCA}$ and $\hat{L}$ are close enough. Since the algorithm acts upon random matrices, we aim for bounds that hold with high probability. In (2.0.13) we already pointed out that

$$\dim \text{span}\{H[f](x) \mid x \in \mathbb{R}^d\} \leq \frac{m(m+1)}{2}. \quad (3.3.6)$$

It is reasonable to assume that the dimension of the span does not change when we restrict $x$ to the unit sphere $x \in \mathbb{S}^{d-1}$. However, the dimension of the span also depends on the rank of $G_x B S_x B^T G_x + T_x \in \mathbb{R}^{m \times m}$ (cf. 2.0.18). To show the concentration, we need to make sure that the rank of $M$ is sufficiently large. Due to (2.0.12), all columns of $M$ lie in the space spanned by $\text{vec}(a_j \otimes a_i + a_i \otimes a_j), 1 \leq i \leq j \leq m$. One assumption could be, that, by sampling enough vectors $\text{vec}(H[f](x_1)), \ldots, \text{vec}(H[f](x_{m_x}))$, we will get the space

$$\text{span}\{\text{vec}(H[f](x_n)) \mid n = 1, \ldots, m_x\} \cong \text{span}\{\text{vec}(a_j \otimes a_i + a_i \otimes a_j) \mid 1 \leq i \leq j \leq m\}.$$ 

This is equivalent to $\text{rank}(M) = m(m+1)/2$ for $m_x$ large enough, or $\sigma_{m(m+1)/2}(M) > 0$. Here $\sigma_k(M)$ denotes the $k$-th singular value of $M$. Note that we have

$$\sigma_k(M) = \sqrt{\sigma_k(M M^T)}, \quad \text{for } k = 1, \ldots, \max\{m_x, d^2\}. \quad (3.3.6)$$

Additionally

$$M M^T = \sum_{i=1}^{m_x} \text{vec}(H[f](x_i))^\otimes 2 \in \mathbb{R}^{d^2 \times d^2}, \quad (3.3.7)$$

\[28\]
which is a sum of independent matrices. Since \( \text{vec}(H[f](x_i)) \) is bounded, \( \text{vec}(H[f](x_i)) \) has finite second moments and by the law of large numbers

\[
\frac{1}{m_x} \sum_{i=1}^{m_x} \text{vec}(H[f](x_i)) \otimes^2 \to \mathbb{E}_{X \sim \mu_{d-1}} \left[ \text{vec}(H[f](X)) \otimes^2 \right]
\]

for \( m_x \to \infty \). Therefore,

\[
\sigma_k(M) \approx \sqrt{m_x \sigma_k \left( \mathbb{E} \left[ \text{vec}(H[f](X)) \otimes^2 \right] \right)}
\]

for \( m_x \) large. The quality of this approximation depends on the number of samples \( m_x \) and the deviation of \( \text{vec}(H[f](X)) \otimes^2 \) from its mean. We want to show that the Hessians concentrate around a space of dimension \( m + m_1 \). If we want to be able to construct the full space, we need \( \left( \mathbb{E}_{\mu_{d-1}} \left[ \text{vec}(H[f](X)) \otimes^2 \right] \right) \) to have at least rank \( m + m_1 \). So for our analysis we formulate the following assumption:

**Assumption 1.** The \( m + m_1 \)-th singular value of

\[
\mathbb{E}_{\mu_{d-1}} \left[ \text{vec}(H[f](X)) \otimes^2 \right] = \int_{\mathbb{R}^{d-1}} \text{vec}(H[f](x)) \otimes^2 d\mu_{d-1}(x)
\]

is well separated from 0, i.e. there is an \( \alpha > 0 \) such that

\[
\sigma_{m+m_1} \left( \mathbb{E}_{\mu_{d-1}} \left[ \text{vec}(H[f](X)) \otimes^2 \right] \right) \geq \alpha.
\]

(3.3.8)

In the remainder of this section we develop a number of bounds for \( \|P_{L_{PCA}} - P_L\|_F \).

**Theorem 3.** Let \( X_1, \ldots, X_{m_x} \) be i.i.d random vectors in \( \mathbb{R}^d \) such that they are bounded \( \|X_i\|_2 \leq 1 \). For \( f \in \mathcal{F}_d(m_1, m) \) construct

\[
M = [\text{vec}(H[f](X_1)) | \ldots | \text{vec}(H[f](X_{m_x}))] \in \mathbb{R}^{d^2 \times m_x},
\]

\[
\hat{M} = [\text{vec}(P_L H[f](X_1)) | \ldots | \text{vec}(H[f](P_L X_{m_x}))] \in \mathbb{R}^{d^2 \times m_x}.
\]

Denote by \( P_{L_{PCA}}, P_L \) the orthogonal projections onto the spaces spanned by the first \( m + m_1 \) left singular vectors of \( M, \hat{M} \). If \( \|O^T X_i\|_2^2 \) is a sub-gaussian random variable such that \( \mathbb{E}[\|O^T X_i\|_2^2] = \mu, \|\|O^T X_i\|_2^2 - \mu\|_{\psi_2} \leq K \) for any orthogonal matrix \( O \in \mathbb{R}^{d \times m} \) and

\[
\sigma_{m+m_1} \left( \mathbb{E} \left[ \text{vec}(H[f](X_1)) \otimes^2 \right] \right) \geq \alpha > 0,
\]

then we get

\[
\|P_{L_{PCA}} - P_L\|_F \leq 2C_1 \sqrt{\frac{(1 + t)\mu}{\sqrt{(1 - s)\alpha}}}
\]

for \( s \in (0, 1), t > 0, \) with probability at least

\[
1 - 2 \exp \left( -\frac{ct^2 m_x \mu}{K^2} \right) - (m + m_1) \exp \left( -\frac{C_2 m_x \alpha}{C_2 m_1} \right),
\]

and constants \( C_1 = (1 + \epsilon_A)^4(1 + 2\epsilon_B)\eta_2 (\kappa_2^2 + 2\kappa_1 \kappa_2), C_2, c > 0 \).
The proof of Theorem 3 is an application of Wedin’s bound (see Appendix Theorem A.2), which bounds \( \|P_{L_{PCA}} - P_L\|_F \) from above by the distance \( \|M - \hat{M}\|_F \) scaled by the spectral gap of \( M \) and \( \hat{M} \) at the \( m + m_1 \)-th singular value. Before we give a proof, we introduce two auxiliary lemmas.

**Lemma 12.** For \( f \in \mathcal{F}_d(m_1, m) \) define

\[
M = [\text{vec}(H[f](x_1)) | \ldots | \text{vec}(H[f](x_{m_1}))] \in \mathbb{R}^{d \times m_1},
\]

\[
\hat{M} = [\text{vec}(P_L H[f](x_1)) | \ldots | \text{vec}(P_L H[f](x_{m_1}))] \in \mathbb{R}^{d \times m_1},
\]

If \( x_1, \ldots, x_m \in \mathbb{R}^d \) such that \( \|x_i\|_2 \leq 1 \) for all \( i = 1, \ldots, m_1 \), then

\[
\|M - \hat{M}\|_F^2 \leq \tilde{C}_1^2 \sum_{i=1}^{m_1} \|A^T x_i\|_2^2,
\]

for \( \tilde{C}_1 = (1 + \epsilon_A)^3(1 + 2\epsilon_B)\eta_2(\kappa_2^2 + 2\kappa_1\kappa_2) > 0 \).

**Proof of lemma 12.** First, note that

\[
\|M - \hat{M}\|_F^2 = \sum_{i=1}^{m_1} \|\text{vec}(H[f](x_i)) - \text{vec}(P_L H[f](x_i))\|_2^2.
\]

Now

\[
\|\text{vec}(H[f](x_i)) - \text{vec}(P_L H[f](x_i))\|_2 = \|\text{vec}(V_{x_i} S_{x_i} V_{x_i}^T) + \text{vec}(P_L V_{x_i} S_{x_i} V_{x_i}^T) - \text{vec}(P_L A T_{x_i} A^T)\|_2 \quad (3.3.9)
\]

\[
= \|\text{vec}(V_{x_i} S_{x_i} V_{x_i}^T) - \text{vec}(P_L V_{x_i} S_{x_i} V_{x_i}^T)\|_2 = \|V_{x_i} S_{x_i} V_{x_i}^T - P_L V_{x_i} S_{x_i} V_{x_i}^T\|_F,
\]

where the last step is due to \( A T_{x_i} A^T \in L \). Set \( \Delta_{x_i} = G_{x_i} - G_0 \) and denote \( V := V_0 = AG_0B \) then \( \text{vec}(V S_{x_i} V^T) \in L \). Since \( P_L \) is an orthogonal projection, we continue with

\[
\|V_{x_i} S_{x_i} V_{x_i}^T - P_L V_{x_i} S_{x_i} V_{x_i}^T\|_F
\]

\[
\leq \|\text{vec}(V_{x_i} S_{x_i} V_{x_i}^T) - \text{vec}(V S_{x_i} V^T)\|_F
\]

\[
= \|V_{x_i} S_{x_i} V_{x_i}^T - V S_{x_i} V^T\|_F
\]

\[
= \|G_{x_i} B S_{x_i} B^T G_{x_i} - G_0 B S_{x_i} B^T G_0\|_F
\]

\[
\leq (1 + \epsilon_A)^2 \|G_{x_i} B S_{x_i} B^T G_{x_i} - G_0 B S_{x_i} B^T G_0\|_F.
\]

\[
\Rightarrow (1 + \epsilon_A)^2 \|G_{x_i} B S_{x_i} B^T G_{x_i} - G_0 B S_{x_i} B^T G_0\|_F.
\]

\[
\Rightarrow (1 + \epsilon_A)^2 \|\Delta_{x_i} B S_{x_i} B^T \Delta_{x_i} + \Delta_{x_i} B S_{x_i} B^T \Delta_{x_i} + \Delta_{x_i} B S_{x_i} B^T G_0 + G_0 B S_{x_i} B^T \Delta_{x_i}\|_F
\]

\[
\leq (1 + \epsilon_A)^2 \left( \|\Delta_{x_i} B S_{x_i} B^T \Delta_{x_i}\|_F + 2\|\Delta_{x_i} B S_{x_i} B^T G_0\|_F \right)
\]

\[
\leq (1 + \epsilon_A)^2 \left( \|\Delta_{x_i}\|_F^2 \|B S_{x_i} B^T\| + 2\|\Delta_{x_i}\|_F \|B S_{x_i} B^T\| \|G_0\| \right)
\]

\[
\leq (1 + \epsilon_A)^2 (1 + 2\epsilon_B) \|S_{x_i}\| \left( \|\Delta_{x_i}\|_F^2 + 2\kappa_1 \|\Delta_{x_i}\|_F \right).
\]
CHAPTER 3. CONCENTRATION OF THE HESSIANS AROUND L

The elements of $S_{x_i}$ are bounded by $\eta_2$, hence $\|S_{x_i}\| \leq \eta_2$ and $g'_i$ is Lipschitz continuous with Lipschitz constant $\kappa_2$ on $[-1, 1]$ for all $i = 1, \ldots, m$, i.e. $|g'_i(t) - g'_i(0)| \leq \kappa_2 |t|$ for all $t \in [-1, 1]$. Hence

$$(1 + \epsilon_A)^2 (1 + 2\epsilon_B) \|S_{x_i}\| \left( \|\Delta_{x_i}\|^2_F + 2\kappa_1 \|\Delta_{x_i}\|_F \right)$$

$$\leq (1 + \epsilon_A)^2 (1 + 2\epsilon_B) \eta_2 \left( \kappa_2^2 \|AT_{x_i}\|_2^2 + 2\kappa_1 \kappa_2 \|AT_{x_i}\|_2 \right)$$

$$\leq (1 + \epsilon_A)^3 (1 + 2\epsilon_B) \eta_2 \left( \kappa_2^2 + 2\kappa_1 \kappa_2 \right) \|AT_{x_i}\|_2,$$

where we used $\|AT_{x_i}\|_2^2 \leq (1 + \epsilon_A) \|Ax_i\|_2$ in the last step. By introducing the constant $\hat{C}_1 := (1 + \epsilon_A)^3 (1 + 2\epsilon_B) \eta_2 \left( \kappa_2^2 + 2\kappa_1 \kappa_2 \right) > 0$, (3.3.10)

we can now bound $\|M - \hat{M}\|_F^2$ by

$$\|M - \hat{M}\|_F^2 \leq \hat{C}_1 \sum_{i=1}^{m_x} \|AT_{x_i}\|_2^2.$$ 

\[\Box\]

Lemma 13. Let $X_1, \ldots, X_{m_x}$ be i.i.d random vectors in $\mathbb{R}^d$, such that $\|X_i\|_2 \leq 1$. For $f \in \mathcal{F}_d(m_1, m)$ define

$$M = [\text{vec}(H[f](X_1)) | \ldots | \text{vec}(H[f](X_{m_x}))] \in \mathbb{R}^{d^2 \times m_x},$$

$$\hat{M} = [PL \text{vec}(H[f](X_1)) | \ldots | PL \text{vec}(H[f](X_{m_x}))] \in \mathbb{R}^{d^2 \times m_x}.$$ 

If $\|OT_{X_i}\|_2^2$ is sub-gaussian such that $\mathbb{E}[\|OT_{X_i}\|_2^2] = \mu$, $\|OT_{X_i}\|_2^2 - \mu\|_{\Psi_2} \leq K$, for any orthogonal matrix $O \in \mathbb{R}^{d \times m}$, then

$$\|M - \hat{M}\|_F \leq C_1 \sqrt{(1 + t)m_x \mu},$$

with probability at least $1 - 2 \exp \left( -\frac{c^2 m_x \mu}{K^2} \right)$ for some constant $c > 0$ and for all $t > 0$, and $C_1 = (1 + \epsilon_A) \hat{C}_1 = (1 + \epsilon_A)^4 (1 + 2\epsilon_B) \eta_2 \left( \kappa_2^2 + 2\kappa_1 \kappa_2 \right)$.

Proof of Lemma 13. Denote by $\hat{A}$ the optimal orthogonal matrix of $A$. Then by Lemma 12

$$\|M - \hat{M}\|_F^2 \leq \hat{C}_1 \sum_{i=1}^{m_x} \|\hat{A}T_{X_i}\|_2^2$$

$$\leq (1 + \epsilon_A)^2 \hat{C}_1 \sum_{i=1}^{m_x} \|\hat{A}T_{X_i}\|_2^2 = C_1 \sum_{i=1}^{m_x} \|\hat{A}T_{X_i}\|_2^2$$

for $C_1 = (1 + \epsilon_A) \hat{C}_1$. Set $Y_i := \|\hat{A}T_{X_i}\|_2^2$. Then by our assumption $\mathbb{E}[Y_i] = \mu$ and $\|Y_i - \mu\|_{\Psi_2} \leq K$. Therefore we get

$$\mathbb{P} \left[ \sum_{i=1}^{m_x} \|\hat{A}T_{X_i}\|_2^2 \geq (1 + t)m_x \mu \right] \leq \mathbb{P} \left[ \sum_{i=1}^{m_x} |Y_i - \mathbb{E}[Y_i]| \geq tm_x \mu \right].$$

31
By Hoeffding’s concentration inequality for sub-gaussian random variables (Appendix Theorem A.1), we get for some constant $c > 0$ and $t > 0$

\[
\mathbb{P}
\left[
\sum_{i=1}^{m_x} |Y_i - \mathbb{E}[Y_i]| \geq tm_x \mu
\right]
\leq 2 \exp\left(-\frac{ct^2 m_x^2 \mu^2}{\sum_{i=1}^{m_x} \|Y_i - \mathbb{E}[Y_i]\|_{\psi_2}^2}\right)
\leq 2 \exp\left(-\frac{ct^2 m_x^2 \mu^2}{K^2}\right).
\]

Thus

\[
\|M - \hat{M}\|_F \leq C_1 \sqrt{(1 + t)m_x \mu},
\]

with probability at least

\[
1 - 2 \exp\left(-\frac{ct^2 m_x^2 \mu^2}{K^2}\right).
\]

**Corollary 2.** If $M, \hat{M}$ are constructed according to Algorithm 1 i.e. the $x_i$ are drawn uniformly from the unit sphere in $\mathbb{R}^d$, then

\[
\|M - \hat{M}\|_F \leq C_1 \sqrt{(1 + t)m_x \frac{m}{d}},
\]

with probability at least

\[
1 - 2 \exp\left(-\frac{C_3 t^2 m_x^2 m^2}{d}\right).
\]

**Proof.** Follows directly from the previous lemma and by $\mathbb{E}[\|\bar{A} X_i\|_2^2] = \frac{m}{d}, \|\bar{A} X_i\|_2^2 - \frac{m}{d} \|\hat{\gamma}_2\|_2^2 \leq \frac{C}{2d}$ for $X_i \sim \text{Unif}(S^{d-1})$, which was shown in section 3.2. All that remains is to set $C_3 = c \cdot C/2$. \qed

**Lemma 14.** Let $X_1, \ldots, X_{m_x}$ be i.i.d. random vectors, such that $\|X_i\|_2 \leq 1$. For $f \in \mathcal{F}_d(m_1, m)$ denote

\[
M = [\text{vec}(H[f](X_1)) | \ldots | \text{vec}(H[f](X_{m_x})]) \in \mathbb{R}^{d^2 \times m_x}.
\]

Assume that

\[
\sigma_{m+m_1}(\mathbb{E}[\text{vec}(H[f](X_1))^{\otimes 2}]) \geq \alpha > 0.
\]

Then the $m + m_1$-th singular value of $M$ is bounded by

\[
\sigma_{m+m_1}(M) \geq \sqrt{(1 - t)m_x \alpha},
\]

with probability at least

\[
1 - (m + m_1) e^{-t^2 \frac{m_x \alpha}{C_2 m_1}}
\]

for any $0 < t \leq 1$ and $C_2 = 2(1 + \epsilon_A)^4 (1 + 2\epsilon_B)^2 (\kappa_1^2 \eta_2 + \kappa_2 \eta_1)^2$. 32
CHAPTER 3. CONCENTRATION OF THE HESSIANS AROUND $L$

**Proof of Lemma 14.** Let $x \in \mathbb{R}^d$, $\|x\|_2 \leq 1$, then

$$\|\text{vec}(H[f](x))\|_2 = \|H[f](x)\|_F$$

$$= \|AG_x BS_x B^T G_x A^T + AT_x A^T\|_F$$

$$\leq (1 + \epsilon_A)^2 \|G_x BS_x B^T G_x\|_F + (1 + 2\epsilon_A)\|T_x\|_F.$$ 

We start with a bound for $\|T_x\|_F$. Recall that we defined $S_x \in \mathbb{R}^{m_1 \times m_1}$, $T_x \in \mathbb{R}^{m \times m}$ in (2.0.15 - 2.0.16). The diagonal elements of $T_x$ were given as $T_{x,ii} = \left[ B^T h'(B^T g(A^T x)) \right]_{j\ell} g''_{ij}(a^T_i x)$.

Let $t_x$ be the vector built from the diagonal elements of $T_x$ and denote by '$\cdot$' the element-wise vector multiplication. Then

$$\|T_x\|_F = \|t_x\|_2,$$

and

$$\|t_x\|_2 = \|Bh'(B^T g(A^T x)) \cdot g''(A^T x)\|_2$$

$$\leq \kappa_2 \|Bh'(B^T g(A^T x))\|_2$$

$$\leq \kappa_2 \|B\| \|h'(B^T g(A^T x))\|_2$$

$$\leq (1 + \epsilon_B)\kappa_2 \|h'(B^T g(A^T x))\|_2.$$

$$= (1 + \epsilon_B)\kappa_2 \sqrt{\sum_{i=1}^{m_1} h'_i (h_i g(A^T x))^2}$$

$$\leq (1 + \epsilon_B)\kappa_2 \eta_1 \sqrt{m_1}.$$ 

The other norm can be estimated by

$$\|G_x BS_x B^T G_x\|_F \leq \|G_x\|^2 \|BS_x B^T\|_F$$

$$\leq (1 + 2\epsilon_B)\kappa_1^2 \|S_x\|_F$$

$$\leq (1 + 2\epsilon_B)\kappa_1^2 \eta_2 \sqrt{m_1}.$$ 

Thus $\|\text{vec}(H[f](x))\|_2 \leq (1 + \epsilon_A)^2 (1 + 2\epsilon_B)(\kappa_1^2 \eta_2 + \kappa_2 \eta_1) \sqrt{m_1}$. Note that

$$\sigma_{m+m_1}(M) = \sqrt{\sigma_{m+m_1}(MM^T)}$$

$$= \sqrt{\sigma_{m+m_1}(\sum_{i=1}^{m_1} \text{vec}(H[f](x_i))^\otimes 2)}.$$ 

(3.3.11)

(3.3.12)

The sum $\sum_{i=1}^{m_1} \text{vec}(H[f](X_i))^\otimes 2$ consists of independent random matrices. The summands $\text{vec}(H[f](X_i))^\otimes 2$ are clearly positive semidefinite:

$$z^T \text{vec}(H[f](X_i))^\otimes 2 z = \left( \text{vec}(H[f](X_i))^T z \right)^2 \geq 0$$ for all $z \in \mathbb{R}^d$. 

33
CHAPTER 3. CONCENTRATION OF THE HESSIANS AROUND $L$

This allows us to use the Chernoff concentration inequality for the inner singular values developed by Gittens and Tropp [5], see Appendix Theorem A.5. We will see that the conditions of the theorem trivialize in our case. Let

$$\mathbb{E}[\text{vec}(H[f](X))^\otimes 2] = W\Lambda W^T$$

be the spectral decomposition of $\mathbb{E}[\text{vec}(H[f](X))^\otimes 2]$. Denote by $w_i$ the $i$-th eigenvector. Since $\text{vec}(H[f](X))^\otimes 2$ is positive semidefinite, the same is true for its expectation, so all eigenvalues are non-negative. Further, we can assume the eigenvalues are ordered by their magnitude:

$$\Lambda_1 \geq \ldots \geq \Lambda_d \geq 0.$$ 

Note that this implies $\Lambda_{m+m_1} = \sigma_{m+m_1}(\mathbb{E}[\text{vec}(H[f](X))^\otimes 2]) \geq \alpha$, we can assume that $\alpha$ is such that we have equality $\Lambda_{m+m_1} = \alpha$. Now, set

$$W_-(w_1|w_2| \ldots |w_{m+m_1}) \in \mathbb{R}^{d^2 \times m+m_1} \text{ and } W_+(w_{m+m_1}| \ldots |w_d) \in \mathbb{R}^{d^2 \times (d^2-m-m_1+1)}.$$ 

Since all $\text{vec}(H[f](X_i))^\otimes 2$ are i.i.d and therefore share the same mean we have

$$\mu_{m+m_1} = \lambda_{\max} \left( \sum_{j=1}^{m_2} W_+^T(\mathbb{E}[\text{vec}(H[f](X_i))^\otimes 2])W_+ \right) = m_2 \alpha,$$

and

$$\mu_{m+m_1} = \lambda_{\min} \left( \sum_{j=1}^{m_2} W_+^T(\mathbb{E}[\text{vec}(H[f](X_i))^\otimes 2])W_- \right) = m_2 \alpha.$$ 

We established that $\|\text{vec}(H[f](X))\|_2$ is bounded, therefore we can set

$$\Psi(W) = \|\text{vec}(H[f](X))^\otimes 2\| \leq \|\text{vec}(H[f](X))^\otimes 2\|_F$$

$$\leq \|\text{vec}(H[f](X))\|_2^2 \leq \frac{1}{2} C_2 m_1,$$

with $C_2 := 2(1+\epsilon_4)^4(1+2\epsilon_2)^2(\kappa_2^2 \eta_2 + \kappa_2 \eta_1)^2$.

By Theorem A.5 (Appendix)

$$\mathbb{P} \left[ \lambda_{m+m_1} \left( \sum_{i=1}^{m_2} \text{vec}(H[f](X_i))^\otimes 2 \right) \leq (1-t)m_2 \alpha \right] \leq (m+m_1)e^{-t^2 \frac{m_2 \alpha}{\epsilon_2^2 m_1}}.$$ 

We used the eigenvalues to stay consistent with the notation from the theorem, but since all involved matrices were symmetric and positive semidefinite, the same statement is true for the $m+m_1$-th singular value of $\sum_{i=1}^{m_2} \text{vec}(H[f](X_i))^\otimes 2$. Therefore, and by (3.3.12), we have

$$\sigma_{m+m_1}(M) \geq \sqrt{(1-t)m_2 \alpha}$$

with probability at least $1 - (m+m_1)e^{-t^2 \frac{m_2 \alpha}{\epsilon_2^2 m_1}}$ for any $0 < t \leq 1$. \(\square\)
Proof of Theorem 3. The proof is an application of the previous lemmas together with Wedin’s bound. Given $M, \hat{M}$ let
\[
M = \begin{pmatrix} U_1 & U_2 \end{pmatrix} \begin{pmatrix} \Sigma_1 & 0 & 0 \\ 0 & \Sigma_2 & 0 \end{pmatrix} \begin{pmatrix} V_1^T \\ V_2^T \end{pmatrix},
\]
\[
\hat{M} = \begin{pmatrix} \hat{U}_1 & \hat{U}_2 \end{pmatrix} \begin{pmatrix} \hat{\Sigma}_1 & 0 & 0 \\ 0 & \hat{\Sigma}_2 & 0 \end{pmatrix} \begin{pmatrix} \hat{V}_1^T \\ \hat{V}_2^T \end{pmatrix},
\]
be the singular value decomposition of $M, \hat{M}$, where $\Sigma_1, \hat{\Sigma}_1 \in \mathbb{R}^{m+m_1 \times m+m_1}$ and so on. Note that since $L$ has dimension $m + m_1$ and $\hat{M}$ is the projection of the columns onto $L$, the rank of $\hat{M}$ can not be greater than $m + m_1$. Therefore $\hat{\Sigma}_2 = 0$. Denote by $P_{L_{PCA}}, P_{\hat{L}}$ the projections onto the spaces spanned by the first $m + m_1$ left singular vectors of $M, \hat{M}$. By Wedin’s bound (Appendix Theorem A.2), we have
\[
\| P_{L_{PCA}} - P_{\hat{L}} \|_F \leq \frac{2 \| M - \hat{M} \|_F}{\tilde{\alpha}},
\]
where $\tilde{\alpha} \geq \min_{k,k'} |\sigma_k(\hat{\Sigma}_2) - \sigma_k(\Sigma_1)|$ and $\tilde{\alpha} \geq \min_k |\sigma_k(\Sigma_1)|$. Since $\|X_i\|_2 \leq 1$ and $\mathbb{E} [\text{vec}(H[f](X_1))^{\otimes 2}] \geq \alpha$, Lemma 14 yields
\[
\sigma_{m+m_1}(M) \geq \sqrt{(1-s)m_x \alpha}
\]
with probability at least $1 - (m + m_1)e^{-s^2m_x \alpha}$. This, together with the fact that $\hat{\Sigma}_2 = 0$, allows us to set $\tilde{\alpha} = \sqrt{(1-s)m_x \alpha}$. By Lemma 13, we have
\[
\| M - \hat{M} \|_F \leq C_1 \sqrt{(1+t)m_x \mu},
\]
with probability at least $1 - 2 \exp \left(-\frac{ct^2 m_x \mu}{K^2} \right)$. Combining these with the union bound for probabilities yields
\[
\| P_{L_{PCA}} - P_{\hat{L}} \|_F \leq \frac{2 C_1 \sqrt{(1+t)m_x \mu}}{\sqrt{(1-s)m_x \alpha}} = 2 C_1 \sqrt{\frac{(1+t)\mu}{(1-s)\alpha}},
\]
with probability at least
\[
1 - 2 \exp \left(-\frac{ct^2 m_x \mu}{K^2} \right) - (m + m_1) \exp \left(-\frac{s^2 m_x \alpha}{C_2 m_1} \right),
\]
for $s \in (0, 1), t > 0$. The constants $C_1, C_2, c > 0$ are taken from the respective lemmas. \hfill \qed
CHAPTER 3. CONCENTRATION OF THE HESSIANS AROUND L

Note the immediate consequence of Theorem 3 on Algorithm 1 is:

**Corollary 3.** Under Assumption 1 (cf. (3.3.8)) Algorithm 1 applied to a function \( f \in \mathcal{F}_d(m_1, m) \) produces \( L_{PCA} \) such that

\[
||P_{L_{PCA}} - P_{L}||_F \leq 2C_1 \sqrt{\frac{(1 + t)m}{(1 - s)\alpha d}}
\]

(3.3.13)

with constant \( C_1 = (1 + \epsilon_A)^4(1 + 2\epsilon_B)\eta_2 (\kappa_2^2 + 2\kappa_1\kappa_2) \) and probability at least

\[
1 - 2 \exp \left( -\frac{C_3 s^2 m_x m^2}{d} \right) - (m + m_1) \exp \left( -\frac{s^2 m_x \alpha}{C_2 m_1} \right)
\]

for \( C_2, C_3 > 0 \) constant and \( 0 < s < 1, t > 0 \).

**Proof.** Follows directly from Theorem 3 together with Corollary 2.

The results above guarantee some bound on \( ||P_{L_{PCA}} - P_{L}||_F \). What remains is to show that \( \hat{L}^{vec} \) and \( L^{vec} \) are close.

**Proposition 1.** Let \( f \in \mathcal{F}_d(m_1, m) \) and denote by \( L^{vec}, \hat{L}, L_{PCA} \) the spaces defined above. If \( ||P_{L_{PCA}} - P_{L}||_F \leq \epsilon \) for \( 0 \leq \epsilon < 1 \), then \( \hat{L} \) and \( L^{vec} \) span the same space.

**Proof.** Assume \( ||P_{L_{PCA}} - P_{L}||_F \leq \epsilon \) for \( 0 \leq \epsilon < 1 \). The space \( \hat{L} \) is constructed by the columns of \( \hat{M} \) which lie in \( L^{vec} \), therefore \( \hat{L} \) has to be a subspace of \( L^{vec} \). By Theorem 1 we know that \( L \) is spanned by \( m + m_1 \) independent elements. Therefore

\[
\dim(L^{vec}) = \dim(L) = m + m_1,
\]

and we have \( \hat{L} = L^{vec} \) if \( \dim(\hat{L}) = m + m_1 \). Assume that \( \dim(\hat{L}) < m + m_1 \). Since \( \dim(L_{PCA}) = m + m_1 \), we can find an element \( x \in L_{PCA} \) that is orthogonal to \( \hat{L} \), w.l.o.g \( \|x\|_2 = 1 \). Now consider the distance of the projections of \( x \):

\[
||P_{L_{PCA}}x - P_{\hat{L}}x||_2 \leq ||P_{L_{PCA}} - P_{\hat{L}}||_F \|x\|_2 \leq \epsilon.
\]

However, since \( x \) lies in \( L_{PCA} \) and is orthogonal to \( P_{\hat{L}} \):

\[
||P_{L_{PCA}}x - P_{\hat{L}}x||_2 = \|x - 0\|_2 = 1.
\]

Thus, \( \dim(\hat{L}) \) can not be smaller than \( m + m_1 \).

To some extent, Theorem 3 already implies that \( L^{vec} = \hat{L} \). If \( f \in \mathcal{F}_d(m_1, m) \) is such that Theorem 3 yields \( ||L_{PCA} - \hat{L}||_F < 1 \), then \( L^{vec} = \hat{L} \) with the same probability.

The following proposition is a slight variation of Corollary 3. It might provide a better bound, especially if \( m_1 \) is much smaller than \( m \).
CHAPTER 3. CONCENTRATION OF THE HESSIANS AROUND L

Theorem 4. Under Assumption 1 (cf. (3.3.8)) Algorithm 1 produces $L_{PCA}$ such that

$$
\|P_{L_{PCA}} - P_L\|_F \leq 2C_1 \sqrt{\frac{(1 + t)m_1}{(1-s)\alpha}} \left( \sqrt{\frac{2\log(m)}{d}} + \frac{1}{d} \right)
$$

(3.3.14)

for $C_1 = (1 + \epsilon_A)^4(1 + 2\epsilon_B)\eta_2 (\kappa_2^2 + 2\kappa_1\kappa_2)$, and with probability at least

$$
1 - 2\exp(-2C_3t^2m_xd^{-1}) - (m + m_1)\exp \left( \frac{-s^2m_x\alpha}{C_2m_1} \right)
$$

for $C_2, C_3 > 0$ constant and $s \in (0, 1)$, $t > 0$.

Proof. The proof is very similar to the previous ones, so we will only point out the modifications and use the theory above where possible. Using the notation from the proof of Lemma 12 we had

$$
\|M - \hat{M}\|^2_F \leq \sum_{i=1}^{m_x} \|\text{vec}(H[f](x_i)) - \text{vec}(P_LH[f](x_i))\|^2_2,
$$

together with an upper bound for the columns given by

$$
\|\text{vec}(H[f](x_i)) - \text{vec}(P_LH[f](x_i))\|_2 \leq (1 + \epsilon_A)^2 \left( \|\Delta_{x_i}BS_{x_i}B^T\|_F + 2\|\Delta_{x_i}BS_{x_i}B^TG_0\|_F \right)
$$

for $\Delta_{x_i} = G_{x_i} - G_0$. By slightly changing the next steps in the proof of Lemma 12, we get

$$
(1 + \epsilon_A)^2 \left( \|\Delta_{x_i}BS_{x_i}B^T\Delta_{x_i}\|_F + 2\|\Delta_{x_i}BS_{x_i}B^TG_0\|_F \right)
\leq (1 + \epsilon_A)^2 \left( \|\Delta_{x_i}\|^2\|BS_{x_i}B^T\|_F + 2\|\Delta_{x_i}\|\|G_0\|\|BS_{x_i}B^T\|_F \right)
\leq (1 + \epsilon_A)^2(1 + 2\epsilon_B) \left( \|\Delta_{x_i}\|^2\|S_{x_i}\|_F + 2\|\Delta_{x_i}\|\|G_0\|\|S_{x_i}\|_F \right)
\leq (1 + \epsilon_A)^2(1 + 2\epsilon_B)\eta_2\sqrt{m_1} \left( \|\Delta_{x_i}\|^2 + 2\kappa_1\|\Delta_{x_i}\| \right).
$$

Instead of $\|\Delta_{x_i}\|_F$ we now have $\sqrt{m_1}\|\Delta_{x_i}\|$. Since $\Delta_{x_i}$ is diagonal, it reduces to

$$
\|\Delta_{x_i}\| = \|g'(A^Tx_i) - g'(0)\|_\infty \leq \kappa_2\|A^Tx_i\|_\infty,
$$

where $\|\cdot\|_\infty$ denotes the maximum norm for vectors. Therefore, one gets a slightly modified version of Lemma 12

$$
\|M - \hat{M}\|^2_F \leq \tilde{C}_1^2m_1 \sum_{i=1}^{m_x} \|A^Tx_i\|^2_\infty.
$$

(3.3.15)
Let $\bar{A}$ be the optimal orthogonal matrix w.r.t. $A$ and denote $Y_i = \|\bar{A}^T X_i\|_\infty^2$, where $X_1, \ldots, X_{m_x} \sim \text{Unif}(S^{d-1})$ independently. Denote $\mu = \mathbb{E}[Y_i]$. Then, by Lemma 11

$$\mu \leq \sqrt{\frac{2\log m}{d}} + \frac{1}{d},$$

as well as $\|Y_i - \mathbb{E}[Y_i]\|_\infty^2 \leq \frac{C}{2d+4}$ for some constant $C > 0$ according to the results in the previous section. Additionally, we have

$$\mu \geq \frac{1}{d},$$

simply because the squared first coordinate of a vector that is drawn uniformly from the unit sphere has mean equal to $1/d$. Applying the arguments in the proof of Lemma 13 to (3.3.15) results in

$$\|M - \hat{M}\|_F \leq C_1 \sqrt{(1 + t)m_1 m_x \mu},$$

with probability at least $1 - 2\exp(-2ct^2 m_x d \mu^2)$, for some positive constant $c > 0$. Using the bounds for both sides of $\mu$ gives

$$\|M - \hat{M}\|_F \leq C_1 \sqrt{(1 + t)m_1 m_x \left(\sqrt{\frac{2\log m}{d}} + \frac{1}{d}\right)},$$

(3.3.16)

with probability at least $1 - 2\exp(-2ct^2 m_x d^{-1})$ since, by $\mu \geq \frac{1}{d}$,

$$1 - 2\exp(-2ct^2 m_x d^{-1}) \leq 1 - 2\exp(-2ct^2 m_x d \mu^2).$$

The following steps are equivalent to the proof of Theorem 3: we use Lemma 14 to receive a bound on the $m + m_1$-th singular value, then apply Wedin’s bound with (3.3.16) in the numerator. This yields

$$\|P_{L_{PCA}} - \hat{P}_L\|_F \leq 2C_1 \sqrt{\frac{(1 + t)m_1}{(1 - s)\alpha} \left(\sqrt{\frac{2\log m}{d}} + \frac{1}{d}\right)},$$

with probability at least

$$1 - 2\exp(-2ct^2 m_x d^{-1}) - (m + m_1) \exp\left(-\frac{s^2 m_x \alpha}{C_2 m_1}\right).$$

Setting $C_3 := c$ finishes the proof. 

\hfill \Box
Proposition 2. Assume \( f \in \mathcal{F}_d(m_1, m) \), such that all univariate functions in its definition are equal:

\[
g_1 = \cdots = g_m = h_1 = \cdots = h_m.
\]

In this proposition we will denote those functions by \( g \). If we apply \( g \) to a vector, it is the vector where \( g \) gets applied to every coordinate (cf. \( \text{2.0.5} \)). Denote by \( \tilde{A} \) the optimal orthogonal matrix of \( A \). Let \( Z \sim \text{Unif}(S^{d-1}) \) and denote by \( z_1, z_2 \) the first two marginals of \( Z \). Further, denote by \( Z_m \) the vector containing the first \( m \) marginals of \( Z \). Set

\[
\delta = \|\text{Cov}(\text{vec}(g'(Z_m)^{\otimes 2}))\| + (\mathbb{E}[g'(z_1)^2] - \mathbb{E}[g'(z_1)g'(z_2)])^2.
\]

For \( M, \hat{M} \) defined in \((3.3.4), (3.3.5)\), \( C = (1 + \epsilon_A)^2\eta_2m_1 \), and the matrices

\[
H := \left(\text{vec}(g'(A^T x_1)^{\otimes 2}), \ldots, \text{vec}(g'(A^T x_{m_2})^{\otimes 2})\right),
\]

\[
\tilde{H} := \left(\text{vec}(g'(\tilde{A}^T x_1)^{\otimes 2}), \ldots, \text{vec}(g'(\tilde{A}^T x_{m_2})^{\otimes 2})\right),
\]

we have

\[
\|M - \hat{M}\| \leq (1 + \epsilon_A)^2\eta_2m_1(\sqrt{tm_z\delta} + \|H - \tilde{H}\|), \tag{3.3.17}
\]

with probability at least \( 1 - m^2\left(\frac{t}{4}\right)^{\frac{tm_{x_{m_2}}}{t}} \) for \( t \geq \epsilon \).

Proof. Again, denote by \( \tilde{A}, \tilde{B} \) the optimal orthogonal matrices of \( A, B \). The columns of \( M - \hat{M} \) are given by

\[
\text{vec}(H[f](x_i)) - \text{vec}(P_L H[f](x_i)),
\]

and all lie in \( L_{\perp}^{\text{vec}} \), the space orthogonal to \( L^{\text{vec}} \) (cf. \((3.3.1)\)). Let \( w \in \mathbb{R}^d \), \( \|w\|_2 = 1 \), such that

\[
\|M - \hat{M}\| = \|MT - \hat{M}T\| = \max_{z \in \mathbb{R}^d, \|z\|_2 = 1} \|(MT - \hat{M}T)z\|_2 = \|(MT - \hat{M}T)w\|_2.
\]

We can assume that \( M - \hat{M} \neq 0 \). Therefore, \( w \) has to lie in \( L_{\perp}^{\text{vec}} \) as well. Denote by \( E \in \mathbb{R}^{d \times m_2} \) any matrix, such that all its columns lie in \( L^{\text{vec}} \). Then

\[
\|M - \hat{M} + E\| = \|MT - \hat{M}T + E^T\| \geq \|(MT - \hat{M}T + E^T)w\|_2 \tag{3.3.18}
\]

\[
= \|(MT - \hat{M}T)w + 0\|_2 = \|M - MT\|.
\]

Hence, we can add any matrix \( E \) with columns in \( L^{\text{vec}} \) to \( M - \hat{M} \), without lowering its spectral norm. Denote by \( \tilde{a}_i \) the orthonormal columns of \( \tilde{A} \). Let \( Z \) be distributed according to \( \mu_{g_{d-1}} \) and denote by \( z_1, z_2 \) the first two marginals of \( Z \). Due to the rotational invariance of \( \mu_{g_{d-1}} \) we have

\[
\mu := \mathbb{E}[g'(z_1)g'(z_2)] = \mathbb{E}[g'(\tilde{a}_1^T Z)g'(\tilde{a}_2^T Z)] \text{ for all } i, j = 1, \ldots, m \text{ with } i \neq j, \tag{3.3.19}
\]

39
and

\[ \mathbb{E}[g'(z_i)^2] = \mathbb{E}[g'(\bar{a}_i^T Z)^2] \]

for all \( i = 1, \ldots, m \).

Define \( E \) as the matrix with columns given by

\[ \text{vec}(P_L H[f](x_i)) + \text{vec} \left( A T_{x_i} A^T + \mu A B S_{x_i} B^T A^T \right) \in L^{vec}. \]

By (3.3.18) we have

\[ \| M - \hat{M} \| \leq \| M - \hat{M} + E \|, \]

and the \( i \)-th column of \( M - \hat{M} + E \) is given by

\[
\begin{align*}
&= \text{vec}(H[f](x_i)) - \text{vec}(A T_{x_i} A^T + \mu A B S_{x_i} B^T A^T) \\
&= \text{vec} \left( g'(0)^2 A G_{x_i} B S_{x_i} B^T G_{x_i} A^T - \mu A B S_{x_i} B^T A^T \right) \\
&= (A \otimes A) \sum_{\ell=1}^{m_1} s_{\ell \ell, x_i} \text{vec} \left( (\text{diag}(g'(A T_{x_i})) b_\ell) \otimes b_{x_i} - \mu b_\ell \otimes b_\ell \right) \\
&= (A \otimes A) \sum_{\ell=1}^{m_1} s_{\ell \ell, x_i} \text{vec} \left( (\text{diag}(b_\ell) g'(A T_{x_i})) \otimes \mu (\text{diag}(b_\ell) 1_m) \otimes \mu \right) \\
&= (A \otimes A) \sum_{\ell=1}^{m_1} s_{\ell \ell, x_i} (\text{diag}(b_\ell) \otimes \text{diag}(b_\ell)) \text{vec} \left( g'(A T_{x_i}) \otimes - \mu 1_m \otimes \mu \right),
\end{align*}
\]

where \( 1_m \) stands for the \( m \)-dimensional vector with all ones, and \( 1_{m \times m_x} \) denotes the \( m \times m_x \) matrix with all ones. For the matrices \( H, \bar{H} \), given by

\[
H := \left( \text{vec}(g'(A T_{x_1}) \otimes 2) \right| \ldots | \text{vec}(g'(A T_{x_{m_x}}) \otimes 2) \right), \\
\bar{H} := \left( \text{vec}(g'(\bar{A} T_{x_1}) \otimes 2) \right| \ldots | \text{vec}(g'(\bar{A} T_{x_{m_x}}) \otimes 2) \right),
\]

this yields

\[
\begin{align*}
\| M - \hat{M} + E \| &\leq \|(A \otimes A) \| \sum_{\ell=1}^{m_1} \| \text{diag}(b_\ell) \otimes \text{diag}(b_\ell) \| \| (H - \mu 1_{m \times m_x}) \text{diag}(s_{x_1, \ell \ell}, \ldots, s_{x_{m_x}, \ell \ell}) \| \\
&\leq (1 + \epsilon_A)^2 \eta_2 m_1 \| H - \mu 1_{m \times m_x} \| \\
&\leq (1 + \epsilon_A)^2 \eta_2 m_1 (\| \bar{H} - \mu 1_{m \times m_x} \| + \| \bar{H} - \bar{H} \|).
\end{align*}
\]

Note that the spectral norm of \( \bar{H} - \mu 1_{m \times m_x} \) is given by its largest singular value. We will follow a similar strategy as in Lemma 14 to estimate \( \sigma_1(\bar{H} - \mu 1_{m \times m_x}) \). The first step
is to express \((\bar{H} - \mu \mathbf{1}_{m^2 \times m_2})(\bar{H} - \mu \mathbf{1}_{m^2 \times m_2})^T\) as a sum of independent matrices, which yields

\[
(\bar{H} - \mu \mathbf{1}_{m^2 \times m_2})(\bar{H} - \mu \mathbf{1}_{m^2 \times m_2})^T = \sum_{i=1}^{m_2} (\text{vec}(g'(\bar{A}^T x_i))^{\otimes 2}) - \mu \mathbf{1}_{m^2}^{\otimes 2}. \tag{3.3.20}
\]

Denote by \(\text{Cov}(\text{vec}(g'(\bar{A}^T Z)^{\otimes 2}))\) the covariance matrix of \(\text{vec}(g'(\bar{A}^T Z)^{\otimes 2})\). The \(k\)th element of \(\text{vec}(g'(\bar{A}^T Z)^{\otimes 2})\) is given by \(g'(\bar{a}_i^T Z)g'(\bar{a}_j^T Z)\) for some \(i, j \in \{1, \ldots, m\}\). The one-to-one mapping of \(k\) to \(i, j\) is induced by our vec-operator. This justifies that we address the elements of \(\text{vec}(g'(\bar{A}^T Z)^{\otimes 2})\) by \(g'(\bar{A}^T Z)_{ij}\) and the elements of matrices like \(\text{Cov}(\text{vec}(g'(\bar{A}^T Z)^{\otimes 2}))\) by \(\text{Cov}(\text{vec}(g'(\bar{A}^T Z)^{\otimes 2}))_{i,j,i',j'}\) for \(i, j, i', j' \in \{1, \ldots, m\}\). We have

\[
\mathbb{E} \left[ \text{vec}(g'(\bar{A}^T Z)^{\otimes 2}) - \mu \mathbf{1}_{m^2}^{\otimes 2} \right]_{i,j,i',j'} = \mathbb{E} \left[ g'(\bar{a}_i Z)g'(\bar{a}_j Z)g'(\bar{a}_{i'} Z)g'(\bar{a}_{j'} Z) - g'(\bar{a}_i^T Z)g'(\bar{a}_j^T Z)\mu - g'(\bar{a}_{i'}^T Z)g'(\bar{a}_{j'}^T Z)\mu + \mu^2 \right].
\]

Due to our definition of \(\mu\) in \([3.3.19]\) we can distinct the following cases:

- **Case** \(i \neq j\):

  \[
  \mathbb{E} \left[ g'(\bar{a}_i^T Z)g'(\bar{a}_j^T Z)g'(\bar{a}_{i'}^T Z)g'(\bar{a}_{j'}^T Z) - g'(\bar{a}_i^T Z)g'(\bar{a}_j^T Z)\mu - g'(\bar{a}_{i'}^T Z)g'(\bar{a}_{j'}^T Z)\mu + \mu^2 \right] \\
  = \mathbb{E} \left[ g'(\bar{a}_i^T Z)g'(\bar{a}_j^T Z)g'(\bar{a}_{i'}^T Z)g'(\bar{a}_{j'}^T Z) \right] - \mu^2 - \mu \mathbb{E} \left[ g'(\bar{a}_i^T Z)g'(\bar{a}_{i'}^T Z) \right] + \mu^2 \\
  = \mathbb{E} \left[ g'(\bar{a}_i^T Z)g'(\bar{a}_j^T Z)g'(\bar{a}_{i'}^T Z)g'(\bar{a}_{j'}^T Z) \right] - \mathbb{E} \left[ g'(\bar{a}_i^T Z)g'(\bar{a}_{j'}^T Z) \right] \mathbb{E} \left[ g'(\bar{a}_{i'}^T Z)g'(\bar{a}_j^T Z) \right] \\
  = \text{Cov}(\text{vec}(g'(\bar{A}^T Z)^{\otimes 2}))_{i,j,i',j'}.
\]

- **Case** \(i = j, i' = j'\):

  \[
  \mathbb{E} \left[ g'(\bar{a}_i Z)g'(\bar{a}_j^T Z)g'(\bar{a}_{i'}^T Z)g'(\bar{a}_{j'}^T Z) - g'(\bar{a}_i Z)g'(\bar{a}_j^T Z)\mu - g'(\bar{a}_{i'} Z)g'(\bar{a}_{j'}^T Z)\mu + \mu^2 \right] \\
  = \mathbb{E} \left[ g'(\bar{a}_i Z)g'(\bar{a}_j^T Z) \right]^4 - 2\mathbb{E} \left[ g'(\bar{a}_i Z)g'(\bar{a}_j^T Z) \right] \mu + \mu^2 \\
  = \text{Cov}(\text{vec}(g'(\bar{A}^T Z)^{\otimes 2}))_{i,j,i',j'} + (\mathbb{E} \left[ g'(\bar{a}_i Z)g'(\bar{a}_j^T Z) \right] - \mu)^2.
\]

The case \(i' \neq j'\) is equivalent to \(i \neq j\) due to symmetry. Hence, we can express the expectation as

\[
\mathbb{E} \left[ \text{vec}(g'(\bar{A}^T Z)^{\otimes 2}) - \mu \mathbf{1}_{m^2}^{\otimes 2} \right] = \text{Cov}(\text{vec}(g'(\bar{A}^T Z)^{\otimes 2})) + (\mathbb{E} \left[ g'(\bar{a}_i^T Z)^2 \right] - \mu)^2 D,
\]

where \(D\) is a diagonal matrix with zeros on the diagonal, except for the cases that correspond to \(i = j = i' = j'\). Therefore

\[
\sigma_1(\mathbb{E} \left[ \text{vec}(g'(\bar{A}^T x_i))^{\otimes 2} \right]) = \sigma_1(\mathbb{E} \left[ \text{vec}(g'(\bar{A}^T Z)^{\otimes 2}) \right]) \\
\leq \| \text{Cov}(\text{vec}(g'(\bar{A}^T Z)^{\otimes 2})) \| + \| (\mathbb{E} \left[ g'(\bar{a}_i^T Z)^2 \right] - \mu)^2 D \| \\
= \| \text{Cov}(\text{vec}(g'(Z_m)^{\otimes 2})) \| + (\mathbb{E} \left[ g'(z_1)^2 \right] - \mu)^2,
\]

41
where we use the rotational invariance of $\mu_{d-1}$ in the last step, and $Z_m$ denotes the vector of the first $m$ elements of $Z$. The function $g'$ is bounded by $\kappa_1$ which implies $|\mu| = |E[g'(z_1)g'(z_2)]| \leq \kappa_1^2$. Hence, the matrices in the sum (3.3.20) are bounded:

$$
\| \text{vec}(g'(\bar{A}^T x_i)^{\otimes 2}) - \mu 1_m \|_{\otimes 2} \leq \| \text{vec}(g'(\bar{A}^T x_i)^{\otimes 2}) - \mu \|_{2}^2 = \sum_{i,j=1}^{m} \left( g'((\bar{a}_i x) g'((\bar{a}_j x) - \mu) \right)^2 \leq 4\kappa_1^4 m^2.
$$

Denote $\delta := \| \text{Cov}(\text{vec}(g'(Z_m)^{\otimes 2}))\| + (E[g'(z_1)^2] - \mu)^2$. By the matrix Chernoff inequality (Appendix Theorem A.4), we have for every $t \geq e$:

$$
\| (\bar{H} - \mu 1_{m^2 \times m_x})(\bar{H} - \mu 1_{m^2 \times m_x})^T \| \leq tm_x \delta,
$$

with probability at least $1 - m^2 \left( \frac{e}{t} \right)^{m_x^2 / m^2}$. Hence

$$
\| M - \hat{M} \| \leq \| M - \hat{M} + E \| \leq (1 + \epsilon_A)^2 \eta_2 m_1 (\sqrt{tm_x \delta} + \| H - \bar{H} \|),
$$

with the same probability. \hfill \Box
Chapter 4

Approximation of $L$ by finite differences

4.1 Approximating $\nabla f, H[f]$ by finite differences

At the beginning of the last section we already pointed out that the Hessians $H[f](x)$ are in general not available. Therefore, we have to factor in the error caused by approximating $H[f](x)$. The approximation is done by finite differences. Denote by $e_i$ the $i$-th euclidean standard vector in $\mathbb{R}^d$. The approximation of the gradients $\nabla f(x)$ via finite differences is denoted by $\Delta_\epsilon[f](x)$, where $\epsilon > 0$ is usually very small and

$$\Delta_\epsilon[f](x)_i := \frac{f(x + \epsilon e_i) - f(x)}{\epsilon} \quad \text{for} \quad i = 1, \ldots, d.$$  

Similarly, we denote by $\Delta^2_\epsilon[f](x)$ the approximation of $H[f](x)$, which is given by

$$\Delta^2_\epsilon[f](x)_{ij} := \frac{f(x + \epsilon e_i + \epsilon e_j) - f(x + \epsilon e_i) - f(x + \epsilon e_j) + f(x)}{\epsilon^2} \quad \text{for} \quad i, j = 1, \ldots, d.$$  

In both scenarios we will refer to $\epsilon > 0$ as the step size of the approximation. The following lemmas give an upper bound for the approximation error for our type of function.

To make the following estimates more understandable, we restate the setting of this chapter. As before, we only consider functions $f \in \mathcal{F}_d(m_1, m)$ that fulfill the properties of Definition 9. We refer to $A, B$ as the matrices with columns $a_1, \ldots, a_m$, $b_1, \ldots, b_{m_1}$, which are taken directly from the representation given in Definition 9. This also implies that $\|a_i\|_2, \|b_i\|_2 \leq 1$. Instead of referencing the elements of these vectors directly, we write them as elements of the matrices $A, B$. For example: $a_{ki}$ is the $k^{th}$ coordinate of the vector $a_i$ (for $i \leq m$). The constants $\kappa_i, \eta_j$ are the ones that were defined in (2.0.3) and (2.0.4).

**Lemma 15.** Let $\epsilon > 0$ and $\|x\|_2 \leq 1$. Set $C_1 = 2\max\{\kappa_1^2\eta_2, \kappa_2\eta_1\} > 0$, then

$$\|\nabla f(x) - \Delta_\epsilon[f](x)\|_2 \leq C_1 \epsilon m_1 m. \quad (4.1.1)$$
Proof. For fixed \( k \in \{1, \ldots, d\} \) we define \( \phi(t) = f(x + te_k) - f(x) \). By the mean value theorem there exists a \( 0 < \xi_k < \epsilon \) such that

\[
\Delta_\epsilon[f](x)_k = \frac{\phi(\epsilon) - \phi(0)}{\epsilon} = \phi'(\xi_k)
= \nabla f(x + \xi_k e_k)^T e_k
= \frac{\partial f}{\partial x_k}(x + \xi_k e_k).
\]

Recall that by (2.0.8) the partial derivatives of \( f \) are

\[
\frac{\partial f}{\partial x_k}(x) = \sum_{i=1}^{m} \sum_{i=1}^{m} b_{i\ell} h_i'((b_{\ell}^T g(A^T x))g_i'(a_i^T x)a_{ki}).
\]

Fix \( k \) for now and set \( \bar{x} = x + \xi_k e_k \). Then

\[
|\nabla f(x)_k - \Delta_\epsilon[f](x)_k| = \left| \frac{\partial f}{\partial x_k}(x) - \frac{\partial f}{\partial x_k}(\bar{x}) \right|
= \sum_{i=1}^{m} \sum_{i=1}^{m} b_{i\ell} \left[ h_i'(b_{\ell}^T g(A^T x))g_i'(a_i^T x) - h_i'(b_{\ell}^T g(A^T \bar{x}))g_i'(a_i^T \bar{x}) \right] a_{ki}.
\]

For the inner part of the sum we get

\[
|h_i'(b_{\ell}^T g(A^T x))g_i'(a_i^T x) - h_i'(b_{\ell}^T g(A^T \bar{x}))g_i'(a_i^T \bar{x})|
\leq |h_i'(b_{\ell}^T g(A^T x))g_i'(a_i^T x)| + |h_i'(b_{\ell}^T g(A^T \bar{x}))g_i'(a_i^T \bar{x})|
+ |h_i'(b_{\ell}^T g(A^T x))g_i'(a_i^T \bar{x}) - h_i'(b_{\ell}^T g(A^T \bar{x}))g_i'(a_i^T \bar{x})|
\leq \eta_1 |g_i'(a_i^T x) - g_i'(a_i^T \bar{x})| + \kappa_1 |h_i'(b_{\ell}^T g(A^T x)) - h_i'(b_{\ell}^T g(A^T \bar{x}))|
\leq \eta_1 \kappa_2 |a_{ki}| + \kappa_1 \eta_2 |b_{\ell}^T g(A^T x) - g(A^T \bar{x})|)
\leq \eta_1 \kappa_2 |a_{ki}| + \kappa_2 \eta_2 |b_{\ell}^T a_{kj}|.\]

Hence, with \( \tilde{C}_4 := \max\{\eta_1 \kappa_2, \kappa_2 \eta_2\} \), we get

\[
\left| \frac{\partial f}{\partial x_k}(x) - \frac{\partial f}{\partial x_k}(x + \xi_k e_k) \right|
\leq \tilde{C}_4 \epsilon \sum_{j=1}^{m} \sum_{j=1}^{m} \left( |a_{ki}| + \sum_{j=1}^{m} |b_{ji} a_{kj}| \right) |b_{ji} a_{ki}|.
\]
By the triangle inequality and the previous step, \( \| \nabla f(x) - \Delta_\epsilon[f](x) \|_2 \) can be bounded by

\[
\| \nabla f(x) - \Delta_\epsilon[f](x) \|_2 \leq C_4 \epsilon \left\{ \sum_{k=1}^{d} \left[ \sum_{\ell=1}^{m_1} \sum_{i=1}^{m} \left( |a_{ki}| + \sum_{j=1}^{m} b_{ij}a_{kj} \right) |b_{i\ell}a_{ki}| \right] \right\}^{\frac{1}{2}}
\leq C_4 \epsilon \left\{ \sum_{k=1}^{d} \left[ \sum_{\ell=1}^{m_1} \sum_{i=1}^{m} |b_{i\ell}|^2 \right] \right\}^{\frac{1}{2}} + C_4 \epsilon \left\{ \sum_{k=1}^{d} \left[ \sum_{\ell=1}^{m_1} \sum_{i=1}^{m} |b_{i\ell}a_{ki}| \right] \right\}^{\frac{1}{2}}
\leq C_4 \epsilon \sum_{\ell=1}^{m_1} \sum_{i=1}^{m} |b_{i\ell}| \left\{ \sum_{k=1}^{d} \left[ \sum_{\ell=1}^{m_1} \sum_{i=1}^{m} |b_{i\ell}a_{ki}| \right] \right\}^{\frac{1}{2}} + C_4 \epsilon \sum_{\ell=1}^{m_1} \sum_{i=1}^{m} |b_{i\ell}a_{ki}| \left\{ \sum_{k=1}^{d} \left[ \sum_{\ell=1}^{m_1} \sum_{i=1}^{m} |b_{i\ell}|^2 \right] \right\}^{\frac{1}{2}}
\leq C_4 \epsilon m_1 \sqrt{m} + C_4 \epsilon m_1 m \leq 2C_4 \epsilon m_1 m.
\]

In the last step we used \( \sum_{i=1}^{m} |b_{i\ell}| = \| b_{\ell} \|_1 \leq \sqrt{m} \| b_{\ell} \|_2 = \sqrt{m} \). Setting \( C_4 = 2C_4 \) finishes the proof. \( \Box \)

**Lemma 16.** Let \( \epsilon > 0, \| x \|_2 \leq 1 \), then

\[
\| H[f](x) - \Delta_\epsilon^2[f](x) \|_F \leq C_5 \epsilon m_1 m^{\frac{3}{2}}
\]

(4.1.2)

for a constant \( C_5 > 0 \).

The proof is basically a repetition of the previous one with slightly longer calculations.

**Proof.** Let \( k, n \in \{1, ..., d\} \) and define \( \phi(t) = f(x + t\epsilon_k + \epsilon\epsilon_n) - f(x + \epsilon\epsilon_k) \), by the mean value theorem, there exist \( 0 < \xi_1, \xi_2 < \epsilon \) such that

\[
(\Delta_\epsilon^2[f](x))_{kn} = \frac{\phi(\epsilon) - \phi(0)}{\epsilon^2} = \frac{\phi'(\xi_1)}{\epsilon}
= \frac{\frac{\partial f}{\partial x_k}(x + \xi_1\epsilon_k + \epsilon\epsilon_n) - \frac{\partial f}{\partial x_k}(x + \xi_1\epsilon_k)}{\epsilon}
= \frac{\partial^2 f}{\partial x_k \partial x_n}(x + \xi_1\epsilon_k + \xi_2\epsilon_n).
\]

Hence, we obtain

\[
\| (H[f](x))_{kn} - (\Delta_\epsilon^2[f](x))_{kn} \| = \left| \frac{\partial^2 f}{\partial x_k \partial x_n}(x) - \frac{\partial^2 f}{\partial x_k \partial x_n}(x + \xi_1\epsilon_k + \xi_2\epsilon_n) \right|.
\]
Assume \( k, n \) to be fixed and denote \( \bar{x} = x + \xi_1 e_k + \xi_2 e_n \). By recalling our definition of \( H[f](x) \) in (2.0.9) follows \( \frac{\partial^2 f}{\partial x_k \partial x_n}(x) = \varphi_1(x) + \varphi_2(x) \), where

\[
\varphi_1(x) := \sum_{l=1}^{m_1} \sum_{i,j=1}^{m} h''_l(b_l^T g(A^T x)) g'_i(a_i^T x) g'_j(a_j^T x) a_{ki} a_{nj} b_{i\ell} b_{j}\ell,
\]
\[
\varphi_2(x) := \sum_{l=1}^{m_1} \sum_{i,j=1}^{m} h''_l(b_l^T g(A^T x)) g''_i(a_i^T x) a_{ki} a_{ni} b_{i\ell}.
\]

Thus

\[
|(H[f](x))_{kn} - (\Delta^2 [f](x))_{kn}| \leq |\varphi_1(x) - \varphi_1(\bar{x})| + |\varphi_2(x) - \varphi_2(\bar{x})|.
\]

As before, we start by applying the Lipschitz continuity to the summands of \( |\varphi_1(x) - \varphi_1(\bar{x})| \):

\[
|h''_l(b_l^T g(A^T x)) g'_i(a_i^T x) g'_j(a_j^T x) - h''_l(b_l^T g(A^T \bar{x})) g'_i(a_i^T \bar{x}) g'_j(a_j^T \bar{x})| \\
\leq |h''_l(b_l^T g(A^T x)) g'_i(a_i^T x) g'_j(a_j^T x) - h''_l(b_l^T g(A^T x)) g'_i(a_i^T \bar{x}) g'_j(a_j^T \bar{x})| \\
+ |h''_l(b_l^T g(A^T x)) g'_i(a_i^T \bar{x}) g'_j(a_j^T \bar{x}) - h''_l(b_l^T g(A^T \bar{x})) g'_i(a_i^T \bar{x}) g'_j(a_j^T \bar{x})| \\
\leq \eta_2 |g'_i(a_i^T x) g'_j(a_j^T x) - g'_i(a_i^T \bar{x}) g'_j(a_j^T \bar{x})| + \kappa_2^2 |h''_l(b_l^T g(A^T \bar{x})) - h''_l(b_l^T g(A^T x))| \\
\leq \eta_2 \left[ |g'_i(a_i^T x) g'_j(a_j^T x) - g'_i(a_i^T \bar{x}) g'_j(a_j^T \bar{x})| + |g'_i(a_i^T \bar{x}) g'_j(a_j^T \bar{x}) - g'_i(a_i^T \bar{x}) g'_j(a_j^T \bar{x})| \right] \\
+ \kappa_2^2 \eta_3 \sum_{l=1}^{m} b_{i\ell} \left( g_l(a_i^T x) - g_l(a_i^T \bar{x}) \right) \\
\leq \eta_2 \left[ \kappa_1 |g'_i(a_i^T x) - g'_i(a_i^T \bar{x})| + \kappa_1 |g'_j(a_j^T x) - g'_j(a_j^T \bar{x})| \right] + \kappa_3^2 \eta_3 \sum_{l=1}^{m} b_{i\ell} a_{i\ell}^T (x - \bar{x}) \\
\leq \eta_2 \kappa_1 \kappa_2 \left[ |a_i^T (x - \bar{x})| + |a_j^T (x - \bar{x})| \right] + \kappa_3^2 \eta_3 \sum_{l=1}^{m} b_{i\ell} a_{i\ell}^T (x - \bar{x}) \\
\leq \eta_2 \kappa_1 \kappa_2 \left[ |\xi_1 a_{ki} + \xi_2 a_{ni}| + |\xi_1 a_{kj} + \xi_2 a_{nj}| \right] + \kappa_3^2 \eta_3 \sum_{l=1}^{m} b_{i\ell} (|a_{ki}| + |a_{nj}|) \\
\leq \eta_2 \kappa_1 \kappa_2 \epsilon \left[ |a_{ki}| + |a_{ni}| + |a_{kj}| + |a_{nj}| + \sum_{l=1}^{m} b_{i\ell} (|a_{ki}| + |a_{nj}|) \right] \\
\leq \tilde{C}_5 \epsilon \left[ |a_{ki}| + |a_{ni}| + |a_{kj}| + |a_{nj}| + \sum_{l=1}^{m} b_{i\ell} (|a_{ki}| + |a_{nj}|) \right],
\]

where \( \tilde{C}_5 = \max \{ \eta_2 \kappa_1 \kappa_2, \kappa_3^2 \eta_3 \} \). Hence,

\[
|\varphi_1(x) - \varphi_1(\bar{x})| \leq \sum_{l=1}^{m_1} \sum_{i,j=1}^{m} \tilde{C}_5 \epsilon \left[ |a_{ki}| + |a_{ni}| + |a_{kj}| + |a_{nj}| + \sum_{l=1}^{m} b_{i\ell} (|a_{ki}| + |a_{nj}|) \right] b_{i\ell} b_{j\ell} a_{ki} a_{nj}.
\]
Now
\[ \sum_{\ell=1}^{m_1} \sum_{i,j=1}^{m} \tilde{C}_5 \epsilon \left[ |a_{ki}| + |a_{ni}| + |a_{kj}| + |a_{nj}| \right] |b_{i\ell}b_{j\ell}|a_{ki}a_{nj} \]
\[ = \sum_{\ell=1}^{m_1} \sum_{i,j=1}^{m} \tilde{C}_5 \epsilon \left[ |a_{ki}^2|a_{nj} + |a_{ni}a_{ki}a_{nj}| + |a_{kj}a_{ki}a_{nj}| + |a_{nj}^2a_{ki}| \right] |b_{i\ell}b_{j\ell}|. \]

Applying the triangle inequality of the Frobenius norm results in
\[ \tilde{C}_5 \epsilon \left( \sum_{k,n=1}^{d} \left[ \sum_{\ell=1}^{m_1} \sum_{i,j=1}^{m} |a_{ki}^2a_{nj}| + |a_{ni}a_{ki}a_{nj}| + |a_{kj}a_{ki}a_{nj}| + |a_{nj}^2a_{ki}| \right] |b_{i\ell}b_{j\ell}| \right)^{1/2} \]
\[ \leq 2\tilde{C}_5 \epsilon \left( \sum_{k,n=1}^{d} \left[ \sum_{\ell=1}^{m_1} \sum_{i,j=1}^{m} |a_{ki}^2|b_{i\ell}b_{j\ell}| \right] \right)^{1/2} + 2\tilde{C}_5 \epsilon \left( \sum_{k,n=1}^{d} \left[ \sum_{\ell=1}^{m_1} \sum_{i,j=1}^{m} |a_{ni}a_{ki}a_{nj}|b_{i\ell}b_{j\ell}| \right] \right)^{1/2} \]
\[ \leq 2\tilde{C}_5 \epsilon \sum_{\ell=1}^{m_1} \sum_{i,j=1}^{m} |b_{i\ell}b_{j\ell}| \left( \sum_{k,n=1}^{d} \left[ |a_{ki}^2a_{nj}| \right] \right)^{1/2} + 2\tilde{C}_5 \epsilon \sum_{\ell=1}^{m_1} \sum_{i,j=1}^{m} |b_{i\ell}b_{j\ell}| \left( \sum_{k,n=1}^{d} \left[ |a_{ni}a_{ki}a_{nj}| \right] \right)^{1/2} \]
\[ \leq 2\tilde{C}_5 \epsilon \sum_{\ell=1}^{m_1} \sum_{i,j=1}^{m} |b_{i\ell}||b_{j\ell}| \left( \sum_{k,n=1}^{d} \left[ |a_{ki}^4| \sum_{n=1}^{d} a_{nj}^2 \right] \right)^{1/2} + \sum_{k,n=1}^{d} \left[ |a_{ni}a_{ki}a_{nj}| \right] \]
\[ \leq 4\tilde{C}_5 \epsilon \sum_{\ell=1}^{m_1} \sum_{i,j=1}^{m} |b_{i\ell}||b_{j\ell}| \left( \sum_{k,n=1}^{d} \left[ |a_{ki}^4| \sum_{n=1}^{d} a_{nj}^2 \right] \right)^{1/2} \]
\[ \leq 4\tilde{C}_5 \epsilon \sum_{\ell=1}^{m_1} \sum_{i,j=1}^{m} |b_{i\ell}||b_{j\ell}| \left( \sum_{k,n=1}^{d} \left[ |a_{ni}a_{ki}a_{nj}| \right] \right) \]
\[ \leq 4\tilde{C}_5 \epsilon \sum_{\ell=1}^{m_1} \sum_{i,j=1}^{m} |b_{i\ell}||b_{j\ell}| \left( \sum_{k,n=1}^{d} \left[ |a_{ni}a_{ki}a_{nj}| \right] \right)^{1/2} \]
\[ \leq 4\tilde{C}_5 \epsilon m_1 m. \]

The last inequalities are due to \( \|a_i\|_2 = 1 \) for all \( i = 1, \ldots, m \) and \( \|b_{i\ell}\|_1 \leq \sqrt{m}\|b_{i\ell}\|_2 = \sqrt{m} \) for all \( \ell = 1, \ldots, m_1 \). A very similar computation yields
\[ \tilde{C}_5 \epsilon \left( \sum_{k,n=1}^{d} \left[ \sum_{\ell=1}^{m_1} \sum_{i,j=1}^{m} (|a_{ki}| + |a_{ni}|) |a_{ki}a_{nj}|b_{i\ell}b_{j\ell} |b_{i\ell}b_{j\ell}| \right] \right)^{1/2} \leq 4\tilde{C}_5 \epsilon m_1 m^{3/2}. \]

Combining both results gives
\[ \left\{ \sum_{k,n=1}^{d} \left[ \varphi_1(x) - \varphi_1(x + \xi_{1,\ell}e_k + \xi_{2,\ell}e_n) \right]^2 \right\}^{1/2} \leq 8\tilde{C}_5 \epsilon m_1 m^{3/2}. \]
Here we denote $\xi_{1, kn}, \xi_{2, kn}$, to make clear that $\xi_1, \xi_2$ are changing for every partial derivative of second order. However, all $\xi_{1, kn}, \xi_{2, kn}$ are bounded by $\epsilon$, so our result still holds. Applying the same procedure to $|\varphi_2(x) - \varphi_2(\bar{x})|$ yields

$$|h_t'(b_t^T g(A^T x))g_t''(a_t^T x) - g_t'(b_t g(A^T \bar{x}))g_t''(a_t^T \bar{x})|$$

$$\leq \eta_1 \kappa_3 (|a_{ki}| + |a_{mi}|) + \kappa_2 \eta_2 \kappa_1 \epsilon \sum_{I=1}^{m_1} b_{It}(|a_{Ik}| + |a_{In}|).$$

By setting $\hat{C}_5 = \max \{\eta_1 \kappa_3, \kappa_1 \kappa_2 \eta_2\}$, we can develop the same bounds for both parts of the right sum as for $\varphi_1$, and get

$$\left\{ \sum_{k, n=1}^{d} \left[ |\varphi_2(x) - \varphi_2(x + \xi_{1, kn} e_k + \xi_{2, kn} e_n)| \right]^2 \right\}^{\frac{1}{2}} \leq 8 \hat{C}_5 \epsilon m_1 m_2^2.$$

Finally, we get

$$\|H[f](x) - \Delta^2[f](x)\|_F \leq \left\{ \sum_{k, n=1}^{d} \left[ \frac{\partial^2 f}{\partial x_k \partial x_n}(x) - \frac{\partial^2 f}{\partial x_k \partial x_n}(x + \xi_{1, kn} e_k + \xi_{2, kn} e_n) \right]^2 \right\}^{\frac{1}{2}}$$

$$= \left\{ \sum_{k, n=1}^{d} \left[ |\varphi_1(x) + \varphi_2(x) - \varphi_1(x + \xi_{1, kn} e_k + \xi_{2, kn} e_n) - \varphi_2(x + \xi_{1, kn} e_k + \xi_{2, kn} e_n)| \right]^2 \right\}^{\frac{1}{2}}$$

$$\leq 8 \hat{C}_5 \epsilon m_1 m_2^2 + 8 \hat{C}_5 \epsilon m_1 m_2^2.$$

Setting $C_5 = 16 \max \{\hat{C}_5, \hat{C}_5\}$ finishes the proof.

The result is interesting because the error does not depend on $d$, further we did not use the quasi-orthogonality of $A, B$ in any of the calculations, the fact that $\|a_i\|_2, \|b_i\|_2 \leq 1$ was sufficient.

### 4.2 Approximation of $L$

The succeeding theorem extends Theorem 4 to finite differences. This extension can also be done with every other bound on $\|P_L - P_{LPCA}\|_F$ from the previous section.

We can easily extend Algorithm 1 to finite differences. Simply replace the matrix

$$M = (\text{vec}(H[f](x_1))) \ldots \text{vec}(H[f](x_{m_x}))$$

by

$$\tilde{M} := \left( \text{vec}(\Delta^2[f](x_1)) \ldots \text{vec}(\Delta^2[f](x_{m_x})) \right),$$

where $\Delta^2[f](x)$ is the finite difference approximation of $H[f](x)$ and $x_1, \ldots, x_{m_x}$ are the vectors that were drawn uniformly from the sphere. The result will be the space $\tilde{L}_{PCA}$.
CHAPTER 4. APPROXIMATION OF $L$ BY FINITE DIFFERENCES

spanned by the first $m + m_1$ columns of the left singular vectors of $\hat{M}$. As previously, we consider the matrix

$$\hat{M} = (\text{vec}(P_L H[f](x_1)) \ldots \text{vec}(P_L H[f](x_{m_2}))),$$

and denote by $\hat{L} \subset L^\text{vec}$ (cf. (3.3.5)) the space spanned by the columns of $\hat{M}$. Let $P_L, P_{L^\text{PCA}}, P_{\hat{L}}$ be orthogonal projections onto the spaces $L, L_{\text{PCA}}, \hat{L}$.

**Theorem 5.** Given the setting above: If Assumption 1 holds for $\alpha > 0$, then Algorithm 1 applied with finite difference approximations yields $L_{\text{PCA}}$ such that, for every $0 < t, s \in (0, s_0]$, we have

$$\|P_{L_{\text{PCA}}} - P_L\|_F \leq \frac{2C_5\epsilon m_1 m^3 + 2C_1 \sqrt{(1 + t)m_1 \left( \sqrt{\frac{2\log(m)}{d}} + \frac{1}{d} \right)}}{\sqrt{(1 - s)\alpha} - C_5\epsilon m_1 m^3},$$

with probability at least $1 - 2 \exp \left(-2C_3 t^2 m x \frac{m}{d}\right) - (m + m_1) \exp \left(-s^2 \frac{m x \alpha}{C_2 m_1}\right)$, where $\epsilon > 0$ is the step size used in the finite difference approximation and $s_0 \in (0, 1)$ is such that

$$\sqrt{(1 - s_0)\alpha} > C_5\epsilon m_1 m^3.$$

The constants $C_1, C_2, C_3, C_5 > 0$ were defined within the previous sections.

**Proof.** By the proof of Theorem 4, we have

$$\|M - \hat{M}\|_F \leq C_1 \sqrt{(1 + t)m x \delta}$$

for

$$\delta := m_1 \left( \sqrt{\frac{2\log(m)}{d}} + \frac{1}{d} \right),$$

with probability at least $1 - 2 \exp \left(-2C_3 t^2 m x \frac{m}{d}\right) (t > 0)$. Combining this with the bound on the approximation error from Lemma 16 yields

$$\|M - \hat{M}\|_F \leq \|M - \hat{M}\|_F + \|M - \hat{M}\|_F$$

$$\leq \left\{ \sum_{i=1}^{m_x} \|H[f](x_i) - \Delta^2[f](x_i)\|_F^2 \right\}^{1/2} + C_1 \sqrt{(1 + t)m x \delta}$$

$$\leq \left\{ \sum_{i=1}^{m_x} \left( C_5\epsilon m_1 m^3 \right)^2 \right\}^{1/2} + C_1 \sqrt{(1 + t)m x \delta}$$

$$\leq \sqrt{m x C_5\epsilon m_1 m^3} + C_1 \sqrt{(1 + t)m x \delta},$$

49
with probability at least $1 - 2 \exp \left(-2C_3 \frac{m_x^2 \alpha}{\delta} \right)$. Since Assumption 1 holds for $\alpha > 0$, we can apply Lemma 14, such that

$$
\sigma_{m+m_1}(M) \geq \sqrt{(1-s)m_x \alpha},
$$

with probability at least $1 - (m + m_1) \exp \left(-s^2 \frac{m_x \alpha}{C_{2m_1}} \right)$. By Weyl’s inequality (see Appendix Theorem A.6) and given that $\sigma_{m+m_1}(M) \geq \sqrt{(1-s)m_x \alpha}$, we get

$$
|\sigma_{m+m_1}(M) - \sigma_{m+m_1}(\tilde{M})| \leq \|M - \tilde{M}\|_F 
\leq \sqrt{m_x C_5 \epsilon m_1 m^2}. \quad (4.2.5)
$$

By our assumption, $\sqrt{(1-s)\alpha} > C_5 \epsilon m_1 m^2$ for all $s < s_0$ (cf. 4.2.4), hence

$$
\sigma_{m+m_1}(\tilde{M}) \geq \sqrt{(1-s)m_x \alpha} - \sqrt{m_x C_5 \epsilon m_1 m^2},
$$

with probability at least $1 - (m + m_1) \exp \left(-s^2 \frac{m_x \alpha}{C_{2m_1}} \right)$ for all $s \in (0, s_0]$. Since the columns of $\tilde{M}$ were projected onto a $m + m_1$ dimensional space, we know it has maximal rank $m + m_1$. Hence, $\sigma_{m+m_1+1}(\tilde{M}) = 0$ and by the same argument as in Proposition 3 of the last chapter, Wedin’s bound gives us

$$
\|P_L - P_{L_{PCA}}\| = \frac{2\|\tilde{M} - \hat{M}\|_F}{|\sigma_{m+m_1}(M) - \sigma_{m+m_1+1}(\tilde{M})|} \leq \frac{2\|\tilde{M} - \hat{M}\|_F}{\sigma_{m+m_1}(M)} 
\leq \frac{2\sqrt{m_x C_5 \epsilon m_1 m^2} + 2C_1 \sqrt{(1+t)m_x \delta}}{\sqrt{(1-s)m_x \alpha} - \sqrt{m_x C_5 \epsilon m_1 m^2}}
\leq \frac{2C_5 \epsilon m_1 m^2 + 2C_1 \sqrt{(1+t)\delta}}{\sqrt{(1-s)\alpha} - C_5 \epsilon m_1 m^2},
$$

with probability at least $1 - 2 \exp \left(-2C_3 \frac{m_x^2 \alpha}{\delta} \right) - (m + m_1) \exp \left(-s^2 \frac{m_x \alpha}{C_{2m_1}} \right)$. \qed

### 4.3 Dimension reduction

The Hessians are symmetric, therefore it will suffice to calculate the lower triangle part of $\Delta^2[f](x)$. However, one still needs $O(m_x d^2)$ function evaluations of $f$ to calculate $\Delta^2[f](x_1), \ldots, \Delta^2[f](x_{m_x})$. To circumvent this behavior, we can apply an approach from [3], which allows us to translate our original problem of learning a function in $\mathbb{R}^d$, to learning a representation of that function in $\mathbb{R}^m$. The reduction is possible since the active subspace of $f \in F_d(m_1, m)$ only has dimension $m$. To be more clear: Let $P_A$ be the orthogonal projection onto the space spanned by $a_1, \ldots, a_m$, then $f(x) = f(P_A x)$. Once
we know \( \text{range}(A) \), we can take any orthonormal basis \( \tilde{A} = (\tilde{a}_1|\ldots|\tilde{a}_m) \) of \( \text{range}(A) \) and get

\[
f(x) = f(P_Ax) = f(\tilde{A}A^T x).
\]

Let \( \tilde{f} : \mathbb{R}^m \rightarrow \mathbb{R} \) be given by \( \tilde{f}(y) = f(\tilde{A}y) \) and note that this implies \( f(x) = \tilde{f}(\tilde{A}^T x) \). By denoting \( \alpha_i = \tilde{A}^T a_i \), the function \( \tilde{f} \) can also be written as

\[
\tilde{f}(y) = \sum_{\ell=1}^{m_1} h_\ell(b_\ell^T \begin{pmatrix}
g_1(\alpha_1^T y) \\
g_2(\alpha_2^T y) \\
\vdots \\
g_m(\alpha_m^T y)
\end{pmatrix}).
\]

Hence \( \tilde{f} \) has the same form as \( f \). For the reduction we need to identify \( \text{span}\{a_1,\ldots,a_m\} \) or a good approximation of it. Recall that, by (2.0.8),

\[
\nabla f(x) \in \text{span}\{a_1,\ldots,a_m\}.
\]

Therefore we can apply the same algorithm as for the Hessians on the gradients. For \( x_1,\ldots,x_{m_x} \) sampled uniformly from the unit sphere in \( \mathbb{R}^d \), denote

\[
A := (\nabla f(x_1)|\ldots|\nabla f(x_{m_x})) \in \mathbb{R}^{d \times m_x}, \quad \text{(4.3.2)}
\]

\[
\tilde{A} := (\Delta_s[f](x_1)|\ldots|\Delta_s[f](x_{m_x})) \in \mathbb{R}^{d \times m_x}. \quad \text{(4.3.3)}
\]

Further denote by \( P_A, P_{A_{PCA}} \) the projection onto the space spanned by the first \( m \) left singular vectors of \( A, \tilde{A} \). We will assume that \( A \) has full rank, therefore \( P_A \) is exactly the projection onto the space \( \text{range}(A) \).

**Proposition 3.** If there exists a \( \beta > 0 \), such that

\[
\sigma_m(\mathbb{E}[(\nabla f(X)(\nabla f(X))^T)]) \geq \beta \quad \text{(4.3.4)}
\]

for \( X \sim \text{Unif}(S^{d-1}) \), then with probability at least \( 1 - m \exp(-\frac{s^2 m_x \beta}{2C_4 \epsilon_m^1}) \), we have

\[
\|P_A - P_{A_{PCA}}\|_F \leq \frac{2C_4 \epsilon m_1 m}{\sqrt{(1-s)\beta - C_4 \epsilon m_1 m}}, \quad \text{(4.3.5)}
\]

for all \( s \in (0, s_0) \), where \( \epsilon > 0 \) is the step size of the finite difference approximation and \( s_0 \in (0,1) \) is such that

\[
\sqrt{(1-s_0)\beta} > C_4 \epsilon m_1 m.
\]
CHAPTER 4. APPROXIMATION OF \( L \) BY FINITE DIFFERENCES

Proof. We start by giving a bound for \( \|\nabla f(x_i)\|_2 \). Let \( x \in \mathbb{R}^d \) such that \( \|x\|_2 \leq 1 \). Then

\[
\|\nabla f(x)\|_2 = \left\| \sum_{\ell=1}^{m_1} \sum_{i=1}^{m} b_{i\ell} h'_{\ell}(b_T \ell g(A^T x)) g'_i(a_T \ell x)a_i \right\|_2
\]

\[
= \left\| \sum_{\ell=1}^{m_1} h'_{\ell}(b_T \ell g(A^T x)) AG_x b_{\ell} \right\|_2 \leq (1 + \epsilon_{AB}) \kappa_1 \|h'(B^T g(A^T x))\|_2 \leq (1 + \epsilon_{AB}) \kappa_1 \eta_1 \sqrt{m_1}.
\]

Denote by \( P = \overline{A}^T \) the orthogonal projector onto the columns of \( \overline{A} \), which denotes the optimal orthogonal matrix of \( A \). Then

\[
\sigma_m(A) = \sigma_m(PA) = \sigma_m(PAA^T P^T) \frac{1}{2} = \left\{ \sum_{i=1}^{m_x} P \nabla f(x_i) \nabla f(x_i)^T P^T \right\}^\frac{1}{2}.
\]

The matrices \( P \nabla f(x_i) \nabla f(x_i)^T P^T \) are clearly symmetric and bounded. Thus,

\[
\|P \nabla f(x_i) \nabla f(x_i)^T P^T\| \leq \|\nabla f(x_i) \nabla f(x_i)^T\|_F \leq \|\nabla f(x_i)\|_2 \leq C_6 m_1,
\]

for \( C_6 = (1 + \epsilon_{AB})^2 \eta_1^2 \kappa_1^2 \). Since all \( x_i \) are independent and from the same distribution, our assumption (4.3.4) yields

\[
\sigma_m(E \sum_{i=1}^{m_x} P \nabla f(x_i) \nabla f(x_i)^T P^T) \geq m_x \beta.
\]

As in the preceding chapter, we can use the matrix Chernoff inequality (Appendix Theorem A.4) to find a lower bound on \( \sigma_m(A) \), which holds with high probability, i.e.

\[
\sigma_m(A) \geq \sqrt{(1 - s) \beta m_x},
\]

with probability at least \( 1 - m \exp(-\frac{s^2 m_x \beta}{2C_6 m_1}) \). Due to Weyl’s theorem (Appendix Theorem A.6), we then have

\[
|\sigma_m(A) - \sigma_m(\tilde{A})| \leq \|A - \tilde{A}\|_F. \tag{4.3.6}
\]

From Lemma 15 follows

\[
\|A - \tilde{A}\|_F = \left( \sum_{i=1}^{m_x} \left\| \nabla f(x_i) - \Delta \epsilon [f](x_i) \right\|_2 \right)^\frac{1}{2} \leq \sqrt{m_x} C_4 \epsilon m_1 m.
\]

If our assumption \( \sqrt{(1 - s_0) \beta} > C_4 \epsilon m_1 m \) holds for \( s_0 \in (0, 1) \), we can combine the results and plug them into (4.3.6):

\[
\sigma_m(\tilde{A}) \geq \sqrt{(1 - s) \beta m_x} - \sqrt{m_x} C_4 \epsilon m_1 m > 0
\]

52
for all $s \in (0, s_0)$. Finally, similar to previous proofs, we can conclude with Wedin’s bound (Appendix Theorem A.2) that

$$\|P_{A_{PCA}} - P_{\tilde{A}_{PCA}}\|_F \leq \frac{\|A - \tilde{A}\|_F}{\sigma_m(\tilde{A}) - \sigma_{m+1}(A)} \leq \frac{\|A - \tilde{A}\|_F}{\sigma_m(A)} \leq 2\sqrt{m_1 m_2 C_4 \epsilon m_1 m} \leq \frac{2C_4 \epsilon m_1 m}{\sqrt{(1 - s) m_2 \beta - C_4 \epsilon m_1 m}},$$

with probability at least $1 - m \exp(-\frac{s^2 m_2 \beta}{2C_6 \epsilon m_1})$. The second step is due to the fact that $A$ has dimension $m$ and therefore $\sigma_{m+1}(A) = 0$.

The last proposition implies that we are able to approximate $\text{span}\{a_1, \ldots, a_m\}$ given $\beta$ is large enough. Assume that $\|P_A - P_{A_{PCA}}\|_F \leq \delta$ for some small $\delta \in (0, 1)$. We can choose any orthonormal basis $\tilde{A} \in \mathbb{R}^{d \times m}$ of range($A_{PCA}$) and define $\tilde{f} : \mathbb{R}^m \to \mathbb{R}$ point wise as $\tilde{f}(y) = f(\tilde{A}y)$. We can sample from $\tilde{f}$ at $y$ by evaluating $f(\tilde{A}y)$. Let us summarize the whole procedure:

**Reduction to the case $d = m$**

Given a function $f \in \mathcal{F}_d(m_1, m)$ (cf. Definition 9):

1. Draw $m_x$ vectors $x_1, \ldots, x_{m_x} \in \mathbb{R}^d$ from $\text{Unif}(S^{d-1})$ at random.
2. Construct $\tilde{A} = (\Delta_\epsilon[f](x_1)| \ldots |\Delta_\epsilon[f](x_{m_x})) \in \mathbb{R}^{d \times m_x}$.
3. Set $\tilde{A}_{PCA}$ to the space spanned by the first $m$ left singular vectors of $\tilde{A}$.
4. Let $\tilde{A} \in \mathbb{R}^{d \times m}$ be the matrix, such that the columns form an orthonormal basis of $\tilde{A}_{PCA}$. Define the function $\tilde{f} : \mathbb{R}^m \to \mathbb{R}$ as $\tilde{f}(y) = f(\tilde{A}^T y)$.

**Theorem 6.** Let $f \in \mathcal{F}_d(m_1, m)$ and denote by $\tilde{f}$ the function constructed by the procedure above. Denote by $A, B$ the matrices corresponding to $f$ (cf. chapter 3), and by $\epsilon_A, \epsilon_B, \epsilon_{AB} \geq 0$ the constants from Definition 9. If there exists a $\delta \geq 0$ such that

$$\|P_A - P_{A_{PCA}}\|_F \leq \delta$$

and

$$0 \leq \sum_{j=1}^{m_1} b_{ij}^2 < 1 - 10\delta - \left(2\epsilon_A + 2\epsilon_B + (\epsilon_A + \epsilon_{AB} + \epsilon_A\epsilon_{AB}) \frac{1}{1 - \epsilon_B}\right)$$

(4.3.7)
for all \( i = 1, \ldots, m \), then
\[
\tilde{f} \in \mathcal{F}_m(m_1, m).
\]

**Proof.** Since \( f \in \mathcal{F}_d(m_1, m) \), it has a representation
\[
f(x) = \sum_{\ell=1}^{m_1} g(b_\ell^T x).
\]
Hence, \( \tilde{f}(y) = f(\tilde{A}^T y) \) can be represented as
\[
\tilde{f}(y) = \sum_{\ell=1}^{m_1} g(b_\ell^T y).
\]
The functions \( g_1, \ldots, g_m, h_1, \ldots, h_{m_1} \) and the vectors \( b_1, \ldots, b_{m_1} \) do not change, hence they still fulfill all properties given in Definition 9. Due to the orthogonality of \( \tilde{A} \), the length of \( \alpha_i \) is bounded, as
\[
\|\alpha_i\|_2 = \|\tilde{A}^T a_i\|_2 \leq \|\tilde{A}\|\|a_i\|_2 \leq \|a_i\|_2 \leq 1,
\]
for all \( i = 1, \ldots, m \). Denote \( P_{\tilde{A}} = \tilde{A}\tilde{A}^T \) and let \( \bar{a}_1, \ldots, \bar{a}_m \) be the optimal orthonormal basis w.r.t. \( a_1, \ldots, a_m \). We follow the arguments in [4, page 9.] and find
\[
S(\alpha_1, \ldots, \alpha_m) = S(\tilde{A}^T a_1, \ldots, \tilde{A}^T a_m) = S(P_{\tilde{A}} a_1, \ldots, P_{\tilde{A}} a_m)
\]
\[
\leq \left( \sum_{i=1}^{m} \|P_{\tilde{A}} a_i - \bar{a}_i\|_2^2 \right)^{\frac{1}{2}}
\]
\[
\leq \left( \sum_{i=1}^{m} \|P_{\tilde{A}} a_i - \bar{a}_i\|_2^2 \right)^{\frac{1}{2}}
\]
\[
\leq \left( \sum_{i=1}^{m} \|P_{\tilde{A}} a_i - P_{\tilde{A}} \bar{a}_i\|_2^2 \right)^{\frac{1}{2}} + \left( \sum_{i=1}^{m} \|P_{\tilde{A}} \bar{a}_i - P_A \bar{a}_i\|_2^2 \right)^{\frac{1}{2}}
\]
\[
\leq \epsilon_A + \delta.
\]
The condition (4.3.7) implicitly implies \( \delta < \frac{1}{10} \) and \( \epsilon_A < \frac{1}{4} \) (otherwise it could not hold). Thus,
\[
S(\alpha_1, \ldots, \alpha_m) \leq \frac{1}{10} + \frac{1}{4} < 1. \tag{4.3.8}
\]
Let \( A^* \in \mathbb{R}^{m \times m} \) be the matrix with columns \( \alpha_1, \ldots, \alpha_m \). If \( \alpha_1, \ldots, \alpha_m \) were not independent, then \( \text{rank}(A^*) < m \). Hence, the \( m^{th} \) singular value of \( A^* \) has to be 0. By Lemma 4 follows
\[
S(\alpha_1, \ldots, \alpha_m) = \left( \sum_{i=1}^{m} (\sigma_i(A^*) - 1) \right)^{\frac{1}{2}} \geq |\sigma_m(A^*) - 1| \geq 1.
\]
This clearly contradicts (4.3.8). Thus, \( \alpha_1, \ldots, \alpha_m \) are independent. We checked the length and linear independence of \( \alpha_1, \ldots, \alpha_m \), which shows that \( \tilde{f} \) fulfills the properties 1-5 of the function class \( \mathcal{F}_m(m_1, m) \). Denote
\[
\epsilon_{A^*} = S(\alpha_1, \ldots, \alpha_m), \quad \epsilon_{A^*B} = \epsilon_{A^*} + \epsilon_B + \epsilon_{A^*}\epsilon_B.
\]
Then the application of a series of loose inequalities in combination with \( \epsilon_A, \epsilon_B \leq \frac{1}{2} \) yields
\[
\begin{align*}
&\left(2\epsilon_{A^*} + 2\epsilon_B + (\epsilon_{A^*} + \epsilon_{A^*B} + \epsilon_{A^*}\epsilon_B) \frac{1}{1 - \epsilon_B}\right) \\
&\leq \left(2\epsilon_A + 2\delta + 2\epsilon_B + \left(4\delta + \epsilon_A + \epsilon_{AB} + \epsilon_A\epsilon_{AB}\right) \frac{1}{1 - \epsilon_B}\right) \\
&\leq 10\delta + \left(2\epsilon_A + 2\epsilon_B + \left(\epsilon_A + \epsilon_{AB} + \epsilon_A\epsilon_{AB}\right) \frac{1}{1 - \epsilon_B}\right).
\end{align*}
\]
Therefore, under the condition (4.3.7), \( \tilde{f} \) fulfills property 6 of Definition 9 and we conclude that \( f \in \mathcal{F}_m(m_1, m) \).

With Theorem 6 in mind, a small \( \delta \) almost guarantees that the reduced function \( \tilde{f} \) will be in \( \mathcal{F}_m(m_1, m) \). In this case every result from the previous section is applicable to \( \tilde{f} \). Approximations of the directions \( \alpha_1, \alpha_2, \ldots \) of \( \tilde{f} \) can be transformed to approximations of \( a_1, a_2, \ldots \) by a transformation with the matrix \( \tilde{A} \), with an error similar to
\[
\| \tilde{A}\alpha_i - a_i \|_2 = \| \tilde{A}\tilde{A}^T a_i - a_i \|_2 = \| \tilde{A}\tilde{A}^T a_i - P_A a_i \|_2 \\
\leq \| \tilde{A}\tilde{A}^T - P_A \|_F \| a_i \|_2 \leq \delta.
\]
The same works for the vectors corresponding to \( v_1, \ldots, v_{m_1} \) in the reduced case.
Chapter 5

Recovering rank-1 matrices from $L$

So far, our only concern was to approximate the space $L$ as given in (2.0.19) by

$$L = \text{span}\left\{a_1^\otimes 2, \ldots, a_m^\otimes 2, v_1^\otimes 2, \ldots, v_{m_1}^\otimes 2\right\}.$$  

The previous chapters show a way to find an approximation $\hat{L} \approx L^{vec}$ (denoted $\hat{L}_{PCA}$ in the last chapter), which merely relied on function evaluations of $f$. In this chapter we will demonstrate methods that make use of the structure of $\hat{L}$ to approximate the directions $a_1, \ldots, a_m, v_1, \ldots, v_{m_1}$. To avoid repetitions we further denote

$$D_A := \{a_1, \ldots, a_m\}, \quad (5.0.1)$$
$$D_V := \{v_1, \ldots, v_{m_1}\}, \quad (5.0.2)$$
$$D := D_A \cup D_V \quad (5.0.3)$$

and similarly

$$D_A^2 := \{a_1^\otimes 2, \ldots, a_m^\otimes 2\}, \quad (5.0.4)$$
$$D_V^2 := \{v_1^\otimes 2, \ldots, v_{m_1}^\otimes 2\}, \quad (5.0.5)$$
$$D^2 := D_A^2 \cup D_V^2. \quad (5.0.6)$$

If not stated otherwise, the directions in $D$ have to be seen in the context of chapter 2. This means they fulfill all the properties inherited by the definition of $F_d(m_1, m)$. As a reminder, the vectors $v_\ell$ were defined as $v_\ell = AG_0 b_\ell$. Throughout this section we will assume that all vectors in $D$ have the length one. We justify this by the fact that changing the length of those elements will not change the space $L$.

5.1 Finding a low-rank basis in a matrix subspace

The first step is to extract approximations of the rank one matrices in $D^2$ from the space $\hat{L}$. We will demonstrate a method and algorithm from [4, Section 3] which addresses this problem. Since we do not add any theory to their approach, we will focus on their
CHAPTER 5. RECOVERING RANK-1 MATRICES FROM L

intuition behind the algorithm and refer the reader to [4] for the theoretical results. The common property of \( a_1 \otimes v_1^\otimes, \ldots, a_m \otimes v_m^\otimes \) is that they are the only matrices of rank one in \( \tilde{L} \). W.l.o.g we can assume that \( \|a_1^\otimes\|_F = \|v_1^\otimes\|_F = 1 \). Therefore, they are the solutions of the nonlinear program

\[
\text{arg min } \text{rank}(M) \text{ s.t. } M \in L, \|M\|_F = 1.
\]

(5.1.1)

Finding the elements of minimal rank in a matrix subspace is by no means trivial. In fact, solving (5.1.1) for a general matrix subspace is NP-hard [12], [1]. Additionally, we can not expect, that solving (5.1.1) for the approximation \( \tilde{L} \) will lead to the desired results, even if the spaces \( L, \tilde{L} \) are very close. However, the fact that \( L \) is spanned by rank one matrices allows for the following relaxation of (5.1.1), given by

\[
\text{arg max } \|M\| \text{ s.t. } M \in L, \|M\|_F = 1,
\]

(5.1.2)

without changing the set of solutions. Simply, because \( \|M\| \leq \|M\|_F \) holds with equality if and only if \( \text{rank}(M) \leq 1 \). The idea is to apply the same NLP to the approximative space and to show that the solutions of

\[
\text{arg max } \|M\| \text{ s.t. } M \in \tilde{L}, \|M\|_F = 1.
\]

(5.1.3)

are close to the rank-1 matrices spanning \( L \). The space studied in [4] is equivalent to \( \text{span}\{a_1 \otimes a_1, \ldots, a_m \otimes a_m\} \). The authors manage to show, that in their case, the local maximizers of (5.1.3) for an approximative space of \( \text{span}\{a_1 \otimes a_1, \ldots, a_m \otimes a_m\} \) are close to the rank-1 matrices \( a_i \otimes a_i \). Their proof is based around the fact that \( a_1, \ldots, a_m \) are quasi-orthogonal. Intuitively, we expect a similar behavior for our case. Especially if the \( b_\ell \) are dense (\( b_i\ell \approx \frac{1}{\sqrt{m}} \)) and \( G_0 \approx I_m \), due to

\[
\langle a_i, v_\ell \rangle = a_i^T A G_0 b_\ell \approx b_i\ell.
\]

By enforcing this, the vectors \( a_1, \ldots, a_m, v_1, \ldots, v_m \) can be made quasi-orthogonal. Unfortunately, we did not manage to show theoretically that for our case the local maximizers of (5.1.3) are close to the matrices \( a_1^\otimes, \ldots, a_m^\otimes, v_1^\otimes, \ldots, v_m^\otimes \). However, our numerical examples suggest that this is the case.

Let \( \gamma > 1 \) and \( X \in \mathbb{R}^{d \times d} \) with singular value decomposition \( X = U \Sigma V^T \). Then \( \Pi_\gamma \) denotes the operator given by

\[
\Pi_\gamma(X) = \frac{1}{\sqrt{\gamma^2 \sigma_1^2 + \sigma_2^2 + \cdots + \sigma_d^2}} U \begin{pmatrix}
\gamma \sigma_1 & 0 & \cdots & 0 \\
0 & \sigma_2 & 0 & \cdots \\
\vdots & \vdots & \ddots & \cdots \\
0 & \cdots & 0 & \sigma_d
\end{pmatrix} V^T.
\]

(5.1.4)

Given an orthogonal projection \( P_\tilde{L} \) that maps into a the matrix space \( \tilde{L} \), we aim to approximate its near-rank-one elements by using the following iterative algorithm from [4].

57
Algorithm 2: Approximating the local maximizers of \((5.1.3)\) [4, cf. Algorithm 5]

\begin{algorithm}
\textbf{Input:} \( P_L \)
\textbf{Parameters:} \( N, \gamma > 1 \)
\textbf{Output:} \( w \)
\begin{algorithmic}[1]
\State \begin{align*}
\text{Choose an initial value } \quad & X^0, \text{ s.t. } \|X^0\|_F = 1; \\
\text{for } & i = 1, \ldots N \ 	ext{do} \\
\quad & \tilde{X}^i \leftarrow P_L \Pi_\gamma (X^{i-1}); \\
\quad & X^i \leftarrow \frac{\tilde{X}^i}{\|X^i\|_F}; \\
\end{align*}
\State \( U\Sigma V^T \leftarrow \text{SVD}(X^N); \)
\State \( w \leftarrow u_1; \quad \text{// } u_1 \text{ denotes the first column of } U \)
\end{algorithmic}
\end{algorithm}

This procedure alternates between increasing the first singular value and projecting the result back onto \( \tilde{L} \). Ideally, it creates a sequence \( X^1, \ldots, X^N \in \tilde{L} \) such that \( \sigma_1(X^1) < \cdots < \sigma_1(X^N) \), which eventually converges, i.e.

\[
\sigma_1(X^N) \approx \max_{M \in \tilde{L}} \|M\|.
\]

At the end it returns the first left singular vector of \( X^N \). If \( X^N \) is close to a symmetric rank-1 matrix \( X^N \approx x \otimes x \), then \( w \approx x \). For an analysis of the convergence we refer to [4]. One crucial part of Algorithm 2 is the choice of the initial \( X^0 \). We expect that Algorithm 2 converges to the local maximum which is closest to \( X^0 \). One reasonable choice is to sample \( X^0 \) randomly in \( \tilde{L} \). In our numerical experiments we take the following approach. Let \( W_1, \ldots, W_k \) be an orthonormal basis of \( \tilde{L} \). Further, let \( \zeta_1, \ldots, \zeta_k \sim \mathcal{N}(0, 1) \). Then we set

\[
\tilde{X}^0 = \sum_{i=1}^k \zeta_i W_i, \quad X^0 = \frac{1}{\|X^0\|_F} \tilde{X}^0. \quad (5.1.5)
\]

5.2 Distinguishing the ridge directions

Let

\[
\Omega := \{ \tilde{a}_1, \ldots, \tilde{a}_m, \tilde{b}_1, \ldots, \tilde{b}_m \} \subset \mathbb{R}^d
\]

be a set of approximations of the vectors in \( \mathcal{D} \), such that every element in \( \Omega \) corresponds to exactly one direction in \( \mathcal{D} \). That means, for distinct \( w_1, w_2 \in \Omega \), we have

\[
\arg \min_{x \in \mathcal{D}} \|w_1 \pm x\|_2 \neq \arg \min_{x \in \mathcal{D}} \|w_2 \pm x\|_2 \quad (5.2.2)
\]
and $|Ω| = m + m_1$. Further, denote by $\epsilon_{opt} > 0$ the minimal $\epsilon > 0$ such that

$$\max_{w \in Ω} \min_{x \in D} \|w \pm x\|_2 \leq \epsilon.$$  \hfill (5.2.3)

Take one $w \in Ω$. How can we distinguish whether $w$ is associated with $D_A$ or $D_V$?

**Problem 1.** Let $Ω$ be a set of approximation such that (5.2.2) holds. Find a decomposition $Ω = Ω_A \cup Ω_V$, $Ω_A \cap Ω_V = \emptyset$, such that

$$\arg \min_{w \in Ω} \|w \pm a\|_2 \in Ω_A \text{ for all } a \in D_A,$$  \hfill (5.2.4)

and

$$\arg \min_{w \in Ω} \|w \pm v\|_2 \in Ω_V \text{ for all } v \in D_V.$$  \hfill (5.2.5)

**Definition 12.** In the notation above, let $Ω$ be a set of approximations. We say the approximations are distinguishable if there exists a decomposition of $Ω$ that solves Problem 1.

**Distinguishing the approximation based on their orthogonality**

To begin with, we start with the simplest case: $A, B$ orthogonal, $G_0 = I_m$ and $Ω = D$, which implies $\epsilon_{opt} = 0$. For $i, j \in \{1, \ldots, m\}$, $\ell, k \in \{1, \ldots, m_1\}$, we have

$$\langle a_i, a_j \rangle = \delta_{ij},$$

$$\langle a_i, v_\ell \rangle = a_i^T A G_0 b_\ell = b_{i\ell},$$

$$\langle v_\ell, v_k \rangle = \delta_{\ell k},$$  \hfill (5.2.6)

where $\delta_{ij}$ denotes the Kronecker delta

$$\delta_{ij} = \begin{cases} 1 & \text{if } i = j \\ 0 & \text{else} \end{cases}.$$  

First, we introduce some terminology from graph theory.

**Definition 13.** Let $G = (V, E)$ be an undirected graph, where $V$ is the set of nodes and $E$ corresponds to the set of edges. An independent set in $G$ is a subset of nodes $C \subset V$ such that every two distinct nodes in $C$ are non-adjacent.

- An independent set is maximal if there is no node $w$ in $V \setminus C$ such that $C \cup \{w\}$ is independent.

- A maximum independent set in $G$ is an independent set $C$ such that there is no other independent set of greater cardinality in $G$. 

59
Proposition 4. In the setting introduced above, the approximations \( \Omega \) are distinguishable if the following criteria are fulfilled:

1. \( m_1 < m \),

2. For any \( 1 \leq k \leq m_1 \) columns of \( B \), the matrix built from those columns has at least \( k + 1 \) non zero rows.

In this case, we can reduce Problem \( \mathcal{P} \) to the maximum independent set problem on a bipartite graph, which is solvable in polynomial time.

Proof. Think of all elements in \( \Omega \) as nodes in a graph \( \mathcal{G} = (W, E) \). Two nodes are connected if and only if the corresponding vectors are not orthogonal. Since the Euclidean product is symmetric, \( \mathcal{G} \) is an undirected graph. Now assume the condition of the proposition is met. By Definition 13 \( C_A = \{a_1, \ldots, a_m\} \) and \( C_V = \{v_1, \ldots, v_{m_1}\} \) are independent sets in \( \mathcal{G} \). This can easily be seen from 5.2.6. Any independent set in \( \mathcal{G} \) corresponds to a set of orthonormal vectors. Since all the vectors lie in the \( m \)-dimensional space \( \text{range}(A) \), an independent set can only contain \( m \) elements. Hence \( C_A \) is a maximum independent set of \( \mathcal{G} \). Due to \( m_1 < m \), \( C_V \) is not a maximum independent set.

Let \( C \neq C_A \) be a maximum independent set in \( \mathcal{G} \). The vectors corresponding to the nodes in \( C \) form an orthonormal basis of \( A \). Denote

\[
n_A := |C \cap C_A|, \quad n_V := |C \cap C_V|.
\]

Then \( 0 < n_A, n_V < m - 1 \). W.l.o.g. we can assume that \( C \) contains the nodes corresponding to the first \( n_A, n_V \) columns of the matrices \( A, V = AG_0B \), i.e.

\[
C = \{a_1, \ldots, a_{n_A}\} \cup \{v_1, \ldots, v_{n_V}\}.
\]

We can construct the matrix \( \mathcal{X} \), given by

\[
\mathcal{X} := (a_1|\ldots|a_{n_A}|v_1|\ldots|v_{n_V}) \in \mathbb{R}^{d \times m}.
\]

Since the columns build an orthonormal basis of \( A \), \( \mathcal{X} \) has to be orthogonal, i.e. \( \mathcal{X}^T \mathcal{X} = I_m \). Write \( I_{m \times n_A} \) for the matrix containing the first \( n_A \) columns of the identity matrix, and \( B_{n_V} \) for the matrix built from the first \( n_V \) columns of \( B \). This allows us to represent \( \mathcal{X} \) as a block matrix

\[
\mathcal{X} = \begin{pmatrix} A I_{m \times n_A} & AG_0 B_{n_V} \end{pmatrix} = A \begin{pmatrix} I_{m \times n_A} & B_{n_V} \end{pmatrix}.
\]

Hence

\[
\mathcal{X}^T \mathcal{X} = \begin{pmatrix} I_{m \times n_A}^T & B_{n_V}^T \\ B_{n_V} & I_{n_V} \end{pmatrix} A^T A \begin{pmatrix} I_{m \times n_A} & B_{n_V} \end{pmatrix}
\]

\[
= \begin{pmatrix} I_{n_A}^T & I_{m \times n_A}^T B_{n_V}^T \\ B_{n_V}^T I_{m \times n_A} & I_{n_V} \end{pmatrix}.
\]
For $X^T X = I_m$ we need $B^T_{n_V} I_{m \times n_A} = 0$. The expression $B^T_{n_V} I_{m \times n_A}$ is exactly the null matrix if the first $n_A$ rows of $B_{n_V}$ are empty, which means that there can only be $m - n_A = n_V$ non-empty rows. This clearly contradicts our assumption that $B_{n_V}$ has at least $n_V + 1$ non-zero rows. So $C$ could not have been a maximum independent set. Hence, $C_A$ is the only maximum independent set in $G$.

If you can split the nodes of a graph into two disjoint sets, such that no edge has endpoints in only one of the sets, the graph is called bipartite. Thus $G$ is bipartite, because the nodes in $C_A$ or $C_V$ are not connected by edges and $C_A \cup C_V = W$. By Theorem A.3 we can find the maximum independent set of a bipartite graph in polynomial time $O((m + m_1)^{2.5} + m)$. This will give us exactly $C_A$. Setting $C_V = W \setminus C_A$ and translating the nodes back to their corresponding vectors finishes the proof.

The general case, in which the directions $a_1, \ldots, a_m, b_1, \ldots, b_{m_1}$ are quasi-orthogonal and in which they were only reconstructed up to some approximation error, is more difficult. Assume that $m_1 < m$, $A, B$ are quasi-orthogonal with $\epsilon_A, \epsilon_B \geq 0$, $G_0 = I_m$, and that we are given approximations

$$\Omega = \{\tilde{a}_1, \ldots, \tilde{a}_m, \tilde{v}_1, \ldots, \tilde{v}_{m_1}\}$$

with maximal error $\epsilon_{\text{opt}} \geq 0$. Then

$$|\langle \tilde{a}_i, \tilde{a}_j \rangle| = |\langle \tilde{a}_i - a_i + a_i, \tilde{a}_j - a_j + a_j \rangle| \leq |\langle \tilde{a}_i - a_i, \tilde{a}_j - a_j \rangle| + |\langle a_i, a_j \rangle| + |\langle a_i, a_j \rangle| \leq \epsilon_{\text{opt}}^4 + 2\epsilon_{\text{opt}}^2 + \epsilon_A,$$

$$|\langle \tilde{a}_i, \tilde{v}_\ell \rangle| \leq \epsilon_{\text{opt}}^4 + 2\epsilon_{\text{opt}}^2 + |\langle a_i, v_\ell \rangle| \leq \epsilon_{\text{opt}}^4 + 2\epsilon_{\text{opt}}^2 + \epsilon_A + \epsilon_B + |b_{i\ell}|,$$

$$|\langle \tilde{v}_\ell, \tilde{v}_k \rangle| \leq \epsilon_{\text{opt}}^4 + 2\epsilon_{\text{opt}}^2 + \epsilon_{AB} = \epsilon_{\text{opt}}^4 + 2\epsilon_{\text{opt}}^2 + \epsilon_A + \epsilon_B + \epsilon_A \epsilon_B.$$

We assume that

$$\epsilon_{\text{opt}}^4 + 2\epsilon_{\text{opt}}^2 + \epsilon_A + \epsilon_B + \epsilon_A \epsilon_B < 1. \quad (5.2.8)$$

Then there exists a $0 \leq \tau_1 \leq 1$ such that

$$\tau_1 \geq \max_{i,j=1,\ldots,m \atop i \neq j} |\langle \tilde{a}_i, \tilde{a}_j \rangle|, \quad \tau_1 \geq \max_{k,\ell=1,\ldots,m_1 \atop k \neq \ell} |\langle \tilde{v}_\ell, \tilde{v}_k \rangle|. \quad (5.2.9)$$

Let $0 \leq \tau_2 < 1$ be such that for any non-empty $K \subset \{1, \ldots, m_1\}$ there is a set $I \subset \{1, \ldots, m\}$ with $|I| \geq |K| + 1$ and

$$\forall i \in I \exists \ell \in K: \tau_2 \leq |\langle \tilde{a}_i, \tilde{v}_\ell \rangle|. \quad (5.2.10)$$
Proposition 5. In the setting above: If we can choose \( \tau_1 < \tau_2 \), we can distinguish the approximations in \( \Omega \) by reducing the Problem to the maximum independent set problem on a bipartite graph, which is solvable in polynomial time.

\[
\text{Proof.} \quad \text{The proof is very similar to the proof of Proposition. Since } \tau_1 < \tau_2, \text{ we can pick } \tau \in (\tau_1, \tau_2).
\]

Let \( C_A, C_V \) be the nodes associated with the approximations \( \{\tilde{a}_1, \ldots, \tilde{a}_m\}, \{\tilde{v}_1, \ldots, \tilde{v}_{m_1}\} \). Due to (5.2.9) and \( \tau > \tau_1 \), \( C_A, C_V \) are independent sets in \( G \). Therefore, we can decompose \( G \) into two independent sets, hence \( G \) is bipartite. Assume \( C \neq C_A \) is a maximum independent set in \( G \) and denote

\[
n_A := |C \cap C_A|, \quad n_V := |C \cap C_V|.
\]

Clearly, \( n_A > 1 \), since \( |C| \geq |C_A| = m > m_1 = |C_V| \geq n_V \). If \( n_V = 0 \), then \( C \subseteq C_A \). In this case \( C \) could only be maximal if \( C = C_A \), which contradicts our assumption that \( C \neq C_A \). Hence \( n_V > 1 \). Let \( K \subseteq \{1, \ldots, m_1\} \) be the set of indices such that the node corresponding to \( \tilde{v}_\ell \) belongs to \( C \) for all \( \ell \in K \). Due to \( \tau < \tau_2 \), we can pick a set \( I \subset \{1, \ldots, m\} \) with \( |I| \geq |K| + 1 = n_V + 1 \), such that

\[
\forall i \in I \exists \ell \in K : \tau < |\langle \tilde{a}_i, \tilde{v}_\ell \rangle|.
\]

This implies that there are at least \( n_V + 1 \) nodes in \( C_A \) which are connected to the nodes in \( C_V \cap C \). Due to the independence of \( C \) those nodes can not be contained in \( C \). Thus \( n_A \leq m - n_V - 1 \), therefore \( |C| = n_A + n_V \leq m - 1 < |C_A| \). This shows that \( C_A \) is the only maximum independent set in \( G \).

By Theorem we can find the maximum independent set of a bipartite graph in polynomial time \( O((m + m_1)^{2.5} + m) \). This will give us exactly \( C_A \). Setting \( C_V = W \setminus C_A \) and translating the nodes back to their corresponding vectors finishes the proof.

Even if the conditions of this approach are fulfilled, it has two major drawbacks:

1. We need to have approximations for all directions in \( D \) before we can distinguish them. In the numerical section we will see that sometimes a few directions in \( D \) are much harder to "find" (i.e. approximate them up to a certain accuracy). To find them at all (or simply to be more efficient) it might be necessary to target those directions specifically. In this case it is helpful if we can distinguish the approximations even if there are some elements missing.

2. The method is not robust against outliers in terms of the approximation error. Suppose a vast majority of our approximations is very accurate, but there is a number of approximations with a high approximation error. Then the behavior of the approach suggested in Proposition becomes completely undefined. We would want a method that is able to distinguish at least the accurate approximations with a high certainty.
Comparing the approximation based on the structure of $f$

We just addressed two problems that occur if we try to distinguish the approximations based on their quasi-orthogonality. These problems make the approach questionable for practical purposes. The structure of $f$ might provide valuable information that helps in solving Problem [1]. In the following section, we sketch a method where $f$ is a function related to neural networks with sigmoidal activation functions. We will focus on the intuition behind the method, which is complemented by numerical results in chapter [6].

In section [1.1] we shortly introduced neural networks and the concept of activation functions. Common examples for sigmoidal activation functions $\phi$ are the hyperbolic tangents or the logistic function. For an illustration of the hyperbolic tangents under slight shifts, see figure [5.2.1]. These functions are typically monotonic increasing and bounded, i.e. there are horizontal asymptotes:

$$\lim_{t\to-\infty} \phi(t) = C_{-\infty}, \quad \lim_{t\to\infty} \phi(t) = C_{\infty}. \quad (5.2.11)$$

In the case $\phi = \tanh$, we have $C_{-\infty} = -1, C_{\infty} = 1$. Additionally, the first derivative $\phi'$ is bell shaped and approaches zero, i.e.

$$\lim_{t\to-\infty} \phi'(t) = 0, \quad \lim_{t\to\infty} \phi'(t) = 0. \quad (5.2.12)$$

Let’s consider the case where $f \in F_d(m_1, m)$, $A, B$ are orthogonal and our approximations are the exact directions $\Omega = D$. As a reminder, the function $f$ has the representation

$$1^T h(B^T g(A^T x)).$$

For the sake of argument we pick $a_1, v_1$ out of $a_1, \ldots, a_m$ and $v_1, \ldots, v_{m_1}$, which are completely arbitrary and could be replaced by any other elements in the corresponding
sets. If we sample $f$ along the direction $a_1$, we get

$$f(ta_1) = 1^T h(B^t g(tA^t a_1)) = 1^T h(B^t g(te_1)),$$

where $e_1$ is the first Euclidean standard vector and $t \in \mathbb{R}$. In the context of neural networks, $g(te_1)$ would be the output (or activation) of the first hidden layer. Suppose that $g_1, \ldots, g_m$ are given by a sigmoidal activation function such that $C := C_\infty = -C_{-\infty}$. Then

$$\lim_{t \to -\infty} g(te_1) = -C \begin{pmatrix} g_2(0) \\ \vdots \\ g_m(0) \end{pmatrix}, \quad \lim_{t \to -\infty} g(te_1) = C \begin{pmatrix} g_2(0) \\ \vdots \\ g_m(0) \end{pmatrix}. \quad (5.2.13)$$

In the terminology of neural nets, we can say that the first neuron is saturated once $|t|$ is sufficiently large, which means that it exhibits values close to $C, -C$. The first layer of the network along $v_1$ is given by

$$g(tA^T AG_0 b_1) = g(tG_0 b_1).$$

Assuming that $G_0 b_1$ is dense (i.e. $b_{i1} \neq 0$ for all $i = 1, \ldots, m$), we get

$$\lim_{t \to -\infty} g(tG_0 b_1) = -\text{sgn}(G_0 b_1) \cdot \begin{pmatrix} C \\ \vdots \\ C \end{pmatrix}, \quad \lim_{t \to -\infty} g(tG_0 b_1) = \text{sgn}(G_0 b_1) \cdot \begin{pmatrix} C \\ \vdots \\ C \end{pmatrix}, \quad (5.2.14)$$

where $\text{sgn}(G_0 b_1) = [\text{sgn}(g_1'(0)b_{11}), \ldots, \text{sgn}(g_m'(0)b_{m1})]^T$ and ‘.’ denotes the element-wise vector multiplication. In the case $(5.2.14)$, every neuron is saturated. Suppose that $t$ is large, such that $(5.2.13)$ and $(5.2.14)$ hold at least approximately. We argue that any small change along a direction of a saturated neuron will not have much impact on the output of the network itself, i.e.

$$g(A^T(ta_1 + \epsilon a_1)) = g((t + \epsilon)A^t a_1) \approx g(tA^t a_1),$$

and since $f$ can be expressed as a continuous function of the first layer, we conclude

$$f(ta_1 + \epsilon a_1) \approx f(ta_1).$$

This implies that

$$\frac{\partial}{\partial a_1} f(ta_1) \approx 0.$$

Along the direction $v_1$, and under the assumption that $b_1$ is dense, the same argument yields

$$\frac{\partial}{\partial a_i} f(tA^T v_1) \approx 0 \quad \text{for all } i = 1, \ldots, m,.$$

64
Since the vectors \( a_1, \ldots, a_m \) form an orthonormal basis of the active subspace of \( f \), the previous equation implies

\[
\| \nabla f(tA^T v_1) \|_2 \approx 0.
\]

On the contrary, we have

\[
\| \nabla f(tA^T a_1) \|_2 = \left( \sum_{i=1}^{m} \left( \frac{\partial}{\partial a_i} f(ta_1) \right)^2 \right)^{\frac{1}{2}} \approx \left( \sum_{i=2}^{m} \left( \frac{\partial}{\partial a_i} f(ta_1) \right)^2 \right)^{\frac{1}{2}},
\]

for the magnitude of the gradient along \( a_1 \). The right side of the last equation is in general constant w.r.t. \( t \) and not 0, since we assumed that \( g'_i(0) \neq 0 \) for all \( i = 1, \ldots, m \) (cf. Definition property 5). By this argument, we can distinguish \( a_1 \) and \( v_1 \) by comparing the magnitude \( \| \nabla f \|_2 \) evaluated along the directions \( a_1, v_1 \). Once we consider the non-orthogonal case or account for the approximation errors of \( a_1 \), the expression \( A^T a_1 \) will most likely be dense as well. However, we argue that \( A^T a_1 \) will still be dominated by its first coordinate, i.e. \( A^T a_1 \approx e_1 \), if the deviation from orthogonality and the approximation error is small enough. Then, the argument above will not hold in the asymptotic case \( |t| \to \infty \), but it will require a very large \( t \) until all neurons are saturated. Hence, we expect that there is an interval \( (t_-, t_+) \), where \( \| \nabla f(tA^T a_1) \|_2 < \| \nabla f(tA^T v_1) \|_2 \). We illustrate this approach in the numerical part (cf. figure 6.2.3 & 6.2.4), in which we account for different degrees of quasi-orthogonality. The setting is given by a neural network with hyperbolic tangents as activation function. The results look very promising and suggest that \( \| \nabla f(tv_\ell) \|_2 \) is decreasing much faster than \( \| \nabla f(ta_i) \|_2 \) for all \( i = 1, \ldots, m, \ell = 1, \ldots, m_1 \).
Chapter 6

Numerical experiment

In the numerical part we primarily study two questions. How many ridge directions can we find, and how large is the final error? The hole approximation scheme consists of several separate steps, each of those steps introduces a potential source of errors. Additionally there is an immense number of possible constellations, this refers to the parameters $m_1, m, d, A, B$ and functions $g_i, h_\ell$. We can only address a few of them, especially since these experiments demand a large number of computations. In section 4.3 it was shown, that one can reduce the general problem to a case where $d = m$. Hence we will only consider cases with $d = m$. Due to our focus on neural networks we only consider functions $g_i, h_\ell$ such that

\[ g_i(x) = \phi_1(x + \theta_i), \quad h_\ell(x) = \phi_2(x + \vartheta_\ell), \]

where $\phi_1, \phi_2$ are activation functions and $\theta_i, \vartheta_\ell \in \mathbb{R}$ represent a bias term. More precisely, we choose one particular representative $\phi_1(x) = \phi_2(x) = \tanh(cx)$ for $c \in (0, 1]$. If not further specified, the shifts $\theta_i, \vartheta_\ell$ are drawn randomly from $\mathcal{N}(0, 0.3)$. Figure 5.2.1 shows the hyperbolic tangents together with its first derivatives under slight shifts $\theta \in [-1, 1]$. We summarize this configuration by defining $\mathcal{F}_{d}(m_1, m)$ as the set of functions $f : \mathbb{R}^d \to \mathbb{R}$ such that:

1. $f \in \mathcal{F}_{d}(m_1, m)$ (cf. Definition 9)

2. The functions $g_i, h_\ell$ have the form

\[ g_i(x) = \tanh(c(x + \theta_i)) \text{ for } i = 1, \ldots, m, \]
\[ h_\ell(x) = \tanh(c(x + \vartheta_\ell)) \text{ for } \ell = 1, \ldots, m_1, \]

where $\theta_1, \ldots, \theta_m, \vartheta_1, \ldots, \vartheta_{m_1} \in \mathbb{R}$.

For our experiments we will consider two slightly modified versions of Algorithm 1, which enable us to sample from spheres with different radii and use more than $m + m_1$ singular vectors (see Algorithm 3 and Algorithm 4). To get a random orthogonal matrix, for
Algorithm 3: Modification of Algorithm 1

Input: $f$
Parameters: $m, k, R$
Output: $L_{PCA}$

begin
1 Sample $x_1, \ldots, x_m$ uniformly from $R \cdot S^{d-1}$;
2 $M \leftarrow \langle \text{vec}(H[f](x_1)) | \ldots | \text{vec}(H[f](x_m)) \rangle$;
3 $U \Sigma V^T \leftarrow \text{SVD}(M)$;
4 $L_{PCA} \leftarrow \text{range}(u_1 | \ldots | u_k)$; // $u_i$ denotes the $i$th column of $U$
end

example $A \in \mathbb{R}^{d \times m}$, we first draw an unitary $d \times d$ matrix from the $O(d)$ Haar distribution\(^1\) and then set $A$ to the first $m$ columns. To sample quasi-orthogonal matrices, we first create an orthogonal matrix, then disturb the singular values by adding a noise vector which is drawn from the multivariate normal distribution. Finally we bring the columns to unit length.

What follows is a template for our experiments and an explanation of the measured quantities. Suppose $f \in \mathcal{F}_m(m_1, m)$, and $L, A, B, a_1, \ldots, a_m, v_1, \ldots, v_{m_1}$ are given by the definitions throughout chapter 2. As previously, denote by

$$
\mathcal{D}_A := \{a_1, \ldots, a_m\}, \\
\mathcal{D}_V := \{v_1, \ldots, v_{m_1}\}, \\
\mathcal{D} := \mathcal{D}_A \cup \mathcal{D}_V.
$$

In the experiments we generally carry out the following steps:

1. Construct a space that approximates $L$, by running Algorithm 3
2. Get $n$ approximations

$$
\Omega = \{w_1, \ldots, w_n\},
$$

by running Algorithm 2 $n$ times.

We measure the following quantities:

- The approximation error for each $w \in \Omega$ up to a sign: $\min_{x \in \mathcal{D}} \| w \pm x \|_2$.
- The average approximation error: $\frac{1}{|\Omega|} \sum_{w \in \Omega} \min_{x \in \mathcal{D}} \| w \pm x \|_2$.

\(^1\)We used the implementation available at http://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.ortho_group.html. See also [14], [11].
• The number of directions in \( \mathcal{D} \) that were recovered (found) up to a certain threshold \( \tau \):

\[
|\{x \in \mathcal{D} \mid \min_{w \in \Omega} \|w \pm x\|_2 \leq \tau\}|.
\]

If all elements were recovered, we speak of a full recovery.

The hole procedure is then repeated for several random configurations of \( f, A, B \), to make the results more reliable.

6.1 Optimal initial values

As mentioned before, the quality of our approximations will depend on the initial values we pick for Algorithm 2. First, we want to answer the question what happens if one picks the optimal initial values. This means we choose \( X^0 = a_1^2, \ldots, a_m^2, v_1^2, \ldots, v_m^2 \) as the initial values of Algorithm 2. The experiments can be seen as a test of the space \( L_{PCA} \).

If \( L_{PCA} \) is not close to \( L_{vec} \), then, regardless of the initial value, the results of Algorithm 2 will be bad. This is due to the fact that Algorithm 2 projects onto the space \( L_{PCA} \).

For the parameters we choose the setting \( d = m = 45, m_1 = 5 \), \( A, B \) orthogonal. The parameter \( k \) in Algorithm 3 indicates the number of singular vectors we use to construct \( L_{PCA} \). Therefore, \( k \) is equivalent to the dimension of \( L_{PCA} \).

6.1.1 Case \( f \in \mathcal{F}^{0.3}_{45}(5, 45) \)

We start with the setting \( f \in \mathcal{F}^{0.3}_{45}(5, 45) \) and \( \epsilon_A, \epsilon_B = 0 \). For each pair \((k, m_x)\), where \( k = 30, 35, \ldots, 125, 130, m_x = 140, 280, 560, \ldots, 35840 \), we repeat the following experiment 15 times.

**Experiment 1: Optimal initial values**

1. Generate \( L_{PCA} \) by using Algorithm 3 with parameters \( k, m_x \) and \( R = 1 \).

2. Generate \( m + m_1 \) approximations by repeating Algorithm 2 \( m + m_1 \) times with parameters \( N = 50, \gamma = 2 \) and initial values \( X^0 = a_1^2, \ldots, a_m^2, v_1^2, \ldots, v_m^2 \).

For each of the 15 repetitions we generate a random configuration by choosing \( A, B \) and \( \theta_i, \varphi_i \sim \mathcal{N}(0, 0.3) \) at random. We measure the number of directions that were recovered up to an error of 0.2 together with their actual approximation error. Figure 6.1.1 shows the percentage of trials where all 50 directions were recovered. While the number of samples seems to have a small effect on the results, the number of singular vectors \( k \) is clearly very significant. We can identify two important transitions starting at \( k = 50 = m + m_1 \) with an accuracy between 15-20\%, which then stagnates until we reach \( k = 95 = 2m + m_1 \). In the last phase after \( k = 2m + m_1 \) we quickly reach full recovery.
While the ratio of full recovery at \( k = m + m_1 \) is rather low, figure 6.1.1 illustrates that for the majority (> 95%) of directions \( k = 50 \) was sufficient. Additionally the right figure indicates that the average error was much lower than 0.2. It seems that there are some directions, which are much less significant in the decomposition of the Hessians, and therefore much harder to reconstruct.

The coefficients of \( a_1^\otimes 2, \ldots, a_m^\otimes 2 \)

The first two graphs in figure 6.1.2 illustrate that the missing directions (i.e. elements that were not recovered up to an error of 0.2) include elements from \( D_A \) as well as \( D_V \) (5.0.3). Particular interesting is that some of the \( a_i \) are not recovered until we reach \( k = 95 \). Recall the definition of the Hessian from (2.0.17)

\[
H[f](x) = \sum_{l=1}^{m_1} s_{x,ll} v_{x,l} \otimes v_{x,l} + \sum_{i=1}^{m} t_{x,ii} a_i \otimes a_i.
\]
The coefficients of \( a_i^{\otimes 2} \) are given by

\[
t_{x,ii} = [B h'(B^T g(A^T x))]_i g''(a_i^T x).
\]

Due to our choice of \( g_i \), the second derivatives \( g''(a_i^T x) \) for different \( i = 1, \ldots, m \) should behave similarly. The remaining part of the coefficients consists of

\[
[B h'(B^T g(A^T x))]_i = b_{i\bullet}' h'(B^T g(A^T x)),
\]

where \( b_{i\bullet} \) denotes the \( i \)-th row vector of \( B \). If \( \|b_{i\bullet}\|_2 \) is small, then the expression (6.1.1) will be small. This is a logical implication inherited by the structure of our function \( f \). Our assumptions on \( B \) do not guarantee any lower bound on the length of its rows.

Assume that \( \|b_{i\bullet}\|_2 = 0 \) for some \( i' \in \{1, \ldots, m\} \). Then \( b_{i\ell} = 0 \) for all \( \ell = 1, \ldots, m_1 \).

Thus

\[
f(x) = 1^T h(B^T g(A^T x)) = \sum_{\ell=1}^{m_1} h_\ell \left( \sum_{i=1}^{m} b_{i\ell} g_i(a_i^T x) \right)
\]

\[
= \sum_{\ell=1}^{m_1} h_\ell \left( \sum_{i=1}^{m} b_{i\ell} g_i(a_i^T x) \right).
\]

The last expression is an exact representation of \( f \) that only uses \( a_1, \ldots, a_{i'-1}, a_{i'+1}, \ldots, a_m \). This touches a very important aspect with respect to our overall problem: Is there a simpler representation of the function \( f \)? By sampling \( B \) at random we will occasionally encounter rows \( b_{i\bullet} \) such that \( \|b_{i\bullet}\|_2 \) is very small compared to the other rows of \( B \). It is reasonable that our algorithm has problems to identify a direction \( a_i \), if it does not have much impact on the function \( f \) itself. The fact that some directions require a much higher number of singular vectors indicates that they are less significant than the noise in our system. If the goal is to approximate \( f \) as good as possible, it might be necessary to find all of the directions. However, if \( f \) represents a neural network with a fixed number of
parameters, then a missing direction might indicate that we can reduce the complexity of the network without losing much of its explainability (at least on the domain we sample from). Answering this question goes beyond the topic of this thesis, but it needs to be acknowledged that an \( a_i \) that can’t be recovered provides valuable information about the structure of the function \( f \).

Regardless of the magnitude of \( b_i \), the coefficient of an \( a_i^2 \) in \( H[f](x) \) will be small if \( b_i \) and \( h'(B^T g(A^T x)) \) are nearly orthogonal. From the previous section we know that we can expect some sort of concentration of the term \( h'(B^T g(A^T x)) \) if we sample \( x \) uniformly from the unit sphere in \( \mathbb{R}^d \). Simply because the marginals of \( A^T x \) concentrate in 0.

In our setting \( h'_t(0) \) will roughly be the same for all \( \ell = 1, \ldots, m_1 \). Since \( \tanh(\theta) \approx 0 \) for \( \theta \) small enough, this will translate to some form of concentration of \( g(A^T x) \approx 0 \). Hence, \( h'(B^T g(A^T x)) \approx h'(0) \approx 1 \). These estimations are very rough, but they still motivate us to check if there is any correlation between \( b_{\ell}^T \mathbf{1} \) and the missing directions \( a_i \). The results can be seen in the histogram of figure 6.1.2. The histogram shows the distribution of \( |\sum_{\ell=1}^{m_1} b_{i\ell}| \) for all \( i = 1, \ldots, m \), where \( i \) is such that \( a_i \) was not recovered. It clearly indicates that for \( k \geq 50 \) the value of \( b_{\ell}^T \mathbf{1} \) is low if \( i \) corresponds to a missing direction. To make this effect more clear we define the family of matrices

\[
\mathcal{B}_t := \{ B \in \mathbb{R}^{m \times m_1} \mid B \text{ orthogonal}, \sum_{\ell=1}^{m_1} b_{i\ell} \geq t \text{ for } i = 1, \ldots, m \}. \quad (6.1.2)
\]

We repeat Experiment 1 another 15 times, sampling \( B \) such that \( B \in \mathcal{B}_{0.04} \). The results are summarized in figure 6.1.3. We can see that the probability of full recovery goes up, especially in the range \( k = 50, \ldots, 95 \). Additionally we can see that our intuition was invalid, by sampling \( B \in \mathcal{B}_{0.04} \) we are able to recover all of the \( a_i \) pretty consistently if \( k \geq m + m_1 = 50 \). Differently put, there seems to be a \( t \) such that Algorithm 3 (with \( k = m + m_1 = 50 \)) recovers all \( a_i \) where \( i \) is such that \( |\sum_{\ell=1}^{m_1} b_{i\ell}| \geq t \).

**Results for different radii of the sphere**

Another important topic we want to address is the choice of distribution, which is used to construct the Hessians. One basic adjustment is to simply change the radius of the sphere. We denote by \( R \) the radius and draw uniformly from \( R \cdot \mathbb{S}^{d-1} := \{ x \in \mathbb{R}^d \mid \|x\|_2 = R \} \).

We repeat Experiment 1 15 times for \( B \in \mathcal{B}_R \) and 15 times for \( B \in \mathcal{B}_{0.04} \). In both settings we keep the number of samples constant at \( m_x = 17920 \) and try different radii \( R = 0.05, 0.5, 1, 2, 5 \). Consulting the analysis section, one might think that increasing the concentration, i.e. lowering the radius of the sphere, leads to better approximations. The results are summarized in figure 6.1.4. Changing the radius has a large impact on the behavior of the algorithm. But the trend is actually the opposite from what we would expect. The performance increases proportional with \( R \). In the case \( B \in \mathcal{B}_{0.04} \) we even reach an accuracy of over 90% for \( k = 50 \) if we draw from the sphere with radius 5. This does not contradict our theory, it merely highlights an important aspect which was not a focus of our analysis. Changing the distribution affects the singular values of the matrix


**Figure 6.1.3:** Experiment 1, $f \in F_{5,45}^{0,3}(5, 45)$, $B \in B_{0.04}$ (cf. page 69): ratio of full recovery (top), average number of recovered directions (bottom left), average number of recovered elements in $\{a_1, \ldots, a_m\}$ (bottom right).
$M$, which is used in Algorithm 3. Optimizing the radius clearly has a positive effect on the overall approximation. However, in a real setting we can not make decision based on the approximation error, since the real directions are unknown. We need to choose the parameters of our distribution, in our case the radius of the sphere $R$, based on the information that is available.

**Gap of the singular values**

One possible criteria is to estimate the quality of a distribution by looking at the singular values $\sigma_1, \sigma_2, \ldots$ of the matrix $M$ (6.1.3). This includes the magnitude of $\sigma_k$ value as well as the gap between $\sigma_k$ and $\sigma_{k+1}$. If the Hessians concentrate around a space of dimension $m + m_1$, our intuition tells us, that there should be drop off between $\sigma_{m+m_1}(M)$ and $\sigma_{m+m_1+1}(M)$. For different values of $1/R \in [0.01, 0.04, 0.07, \ldots, 0.97, 1, 5, 10, \ldots, 95, 100]$ and $m_x = 4480$ we construct

$$M = (\text{vec}(H[f](x_1))|\ldots|\text{vec}(H[f](x_{m_x}))) \in \mathbb{R}^{d^2 \times m_x}$$

(6.1.3)

for 30 random configurations of $A, B, \theta, \vartheta$ and for another 30 random configurations where $B \in \mathcal{B}_{0.04}$. Every time the $x_1, \ldots, x_{m_x}$ are drawn uniformly from $R \cdot \mathbb{S}^{d-1}$. The measured quantities are the magnitude of the $m + m_1$-th singular values as well as the relative gap between $\sigma_{m+m_1}(M)$ and $\sigma_{m+m_1+1}(M)$, which is given by

$$\frac{\sigma_{m+m_1}(M) - \sigma_{m+m_1+1}(M)}{\sigma_{m+m_1}(M)} \in [0, 1].$$

The results are illustrated in figure 6.1.5. The trend in the second row is easy to explain. By increasing the concentration ($R \leq 1$) we reduce the variance of the columns in (6.1.3), therefore we shrink the magnitude of the singular values. Increasing $R$ has exactly the opposite effect. Note that the plots are with respect to the inverse radius $1/R$, which we can be interpreted as the concentration in 0. For $B \in \mathcal{B}_{0.04}$ the gap is increasing with $R$ until it reaches a maximum for $\frac{1}{R} = 0.04$, followed by a decrease at $\frac{1}{R} = 0.01$. This corresponds to $R = 25, 100$. There is a wide range of values between 25 and 100 that were not covered, but it indicates that the maximal (relative) spectral gap lies somewhere between $R = 25$ and $R = 100$. For the general case $B \in \mathcal{B}_0$ the trend is much less significant, there is only a small local maximum at $R = 10$. However, the results in figure 6.1.4 also indicate a increase in performance (in terms of full recovery) for larger radii $R$.

As we argued above, the number of recoverable directions in the general case $B \in \mathcal{B}_0$ is fuzzy, this might be responsible for the absence of a larger spectral gap. For example, if $\|b_1\|_2 \approx 0$, then the gap might be located between $\sigma_{m+m_1-1}$ and $\sigma_{m+m_1}$.
Figure 6.1.4: ratio of full recovery for different radii (cf. page 71): $B \in \mathcal{B}_0$ (top), $B \in \mathcal{B}_{0.04}$ (bottom).
Figure 6.1.5: Plots of the magnitude and the gap of the singular values for different radii of the sphere, with the intention to identify the best radius of the spherical distribution. A description of the setting can be found on page 73. The plots show: magnitude of $\sigma_{m+m_1}$ (middle row), relative gap $\frac{\sigma_{m+m_1} - \sigma_{m+m_1+1}}{\sigma_{m+m_1}}$ (top row), relative gap $\frac{\sigma_m - \sigma_{m+m_1+1}}{\sigma_m}$ (bottom), where $\sigma_k$ denotes the $k^{th}$ singular value of $M$.
Figure 6.1.6: Results for $f \in \mathcal{F}_m^1(m_1, m), m_x = 17920$, with optimal initial values $X^0$ (cf. page 76). The first row show the average ratio of full recovery (left) and the average number of recovered directions (right). The second row show the average approximation error (left), average number of recovered $a_i, v_\ell$ for $k = 50$ (right).

6.1.2 Case $f \in \mathcal{F}_{45}^{1}(5, 45)$

Now we replace the activation function $\tanh(\frac{3x}{10} + \theta)$ by the slightly steeper version $\tanh(x + \theta)$. We distinguish between $B \in B_0, B \in B_{0.04}$ and run Experiment 1 for radii $R \geq 1$ and $k = 50, 95, 130$. The results are summarized in figure 6.1.5. Comparing the ratio of full recovery to the case $f \in \mathcal{F}_{45}^{0.3}(5, 45)$ indicates that increasing the steepness of the activation function decreases the average ratio of full recovery. In contrast to the case $f \in \mathcal{F}_{45}^{0.3}(5, 45)$ we could not find a radius $R$ for which the ratio of full recovery at $k = 50 = m + m_1$ was above 50%. The last plot of figure 6.1.5 shows that increasing $R$ results in a better recovery of the elements $a_1, \ldots, a_m$, however, it has the opposite effect on the recovery of $v_1, \ldots, v_{m_1}$. To avoid this trade-off it might be better to consider a different distribution, where the radius is not fixed.
6.2 Random initial values

Algorithm 4: Modification of Algorithm [1] with finite differences

| Input: f
| Parameters: $m_x, k, R, \epsilon$
| Output: $L_{PCA}$

1 begin
2 Sample $x_1, \ldots, x_{m_x}$ uniformly from $R \cdot \mathbb{S}^{d-1}$;
3 $M \leftarrow \text{vec}(\Delta^2[f](x_1) | \ldots | \Delta^2[f](x_{m_x}))$;
4 $U\Sigma V^T \leftarrow \text{SVD}(M)$;
5 $L_{PCA} \leftarrow \text{range}(u_1 | \ldots | u_k)$; // $u_i$ denotes the $i$th column of $U$
6 end

In this section we will present the results for the general case, where we have to approximate the Hessians by finite differences, pick the initial values $X^0$ randomly and drop the orthogonality assumption. First let us mention some observations. The previous section suggests that increasing the dimension of $L_{PCA}$, denoted by the parameter $k$, leads to significant improvements. We will distinct between the cases $k = m + m_1$ and $k > m + m_1$:

- Case $k = m + m_1$: the average approximation error, and the number of directions we are able to reconstruct up to a certain error were largely determined by the space $L_{PCA}$. Aside from a longer runtime, the error caused by the finite difference approximation was negligible. The same was the case when we replaced the orthogonal matrices by quasi-orthogonal matrices such that $\epsilon_A, \epsilon_B \in (0, 0.3)$. Algorithm 2 with random initial values $X^0 \in L_{PCA}$ (cf. (5.1.5)) performed similar to the algorithm with optimal initial values. The main difference was that it required slightly more iterations until convergence and the ratio of full recovery was less consistent for $f \in \mathcal{F}_{45}^{0.3}(5, 45)$. In summary, the results for the general setting with $k = m + m_1$ came close to the results of the optimal setting.

- Case $k > m + m_1$: We could not reproduce the results of the previous section in the general setting. Algorithm 2 seems to perform poorly if we increase $k$ way beyond $m + m_1$ (for example $k = 2m + m_1$). In almost all cases it was better to choose $k = m + m_1$. 

77
6.2.1 Case $f \in \mathcal{F}_{45}^{0,3}(5, 45)$

For $m_x = 200, 2000, 20000, R = 1, 5, 10, 12, 15, 20$ and different degrees of quasi-orthogonality ($\epsilon_A, \epsilon_B = 0, 0.1, 0.2, 0.5$) we repeat the following experiment 15 times.

**Experiment 2: Random initial values and finite differences**

1. Generate $L_{PCA}$ by using Algorithm 4 with parameters $k = 50$, $m_x$, $R$ and $\epsilon = 0.0001$.

2. Generate $n = 600$ approximations by repeating Algorithm 2 $n$ times with parameters $N = 100, \gamma = 2$ and random initial values $X^0$ sampled according to (5.1.5).

The results are summarized in figure 6.2.1.

6.2.2 Case $f \in \mathcal{F}_{45}^1(5, 45)$

For $m_x = 200, 2000, 20000, R = 1, 10, 13.33, 20$ and different degrees of quasi-orthogonality ($\epsilon_A, \epsilon_B = 0, 0.1, 0.2$) we repeat Experiment 2 15 times. The results are summarized in figure 6.2.2.

**Distinguishing the approximations**

In the last part of section 5.2 we discussed a method to distinguish between the approximations for $a_1, \ldots, a_m$ and $v_1, \ldots, v_{m_1}$. We suggested to make this decision based on the magnitude of the gradient $\|\nabla f\|_2$ along the respective approximations. As an example, we used the results from Experiment 2 in the case $f \in \mathcal{F}_{45}^1(5, 45)$, which was just presented. The parameters were $m_x = 20000, R = 20, B \in \mathcal{B}_{0.04}$. For each case $\epsilon_A, \epsilon_B = 0, 0.1, 0.2$ we recorded 600 approximations. The average approximation error and a histogram of the approximation errors over 15 trials can be seen in the last two rows of figure 6.2.2. For each $x \in \{a_1, \ldots, a_m, v_1, \ldots, v_{m_1}\}$ (cf. figure 6.2.3) and each of the 600 approximations (cf. figure 6.2.4) we plotted $\|\nabla f(tx)\|_2$ for $t \in [-100, 100]$. Each plot shows two regions indicated by the colors blue and red. The blue region contains all the plots of $\|\nabla f(tx)\|_2$ for $x \in \{a_1, \ldots, a_m\}$ (cf. figure 6.2.3) or the corresponding approximations (cf. figure 6.2.4). The red regions contains the plots where $x$ belongs to $\{v_1, \ldots, v_{m_1}\}$ or their corresponding approximations.

In all cases the two regions were well separated. Hence, we can distinguish the approximations according to Definition 12.
Figure 6.2.1: Results of Experiment 2 for $f \in F_{45}^{0.3}(5, 45)$ (cf. page 78). The rows show the ratio of full recovery, the average number of recovered directions, the average approximation error and the histograms (in log-scale) for the approximation errors in the case $R = 20, m_x = 20000$, from top to bottom respectively.
### Figure 6.2.2: Results of Experiment 2 $f \in F_{45}^1(5, 45)$ (cf. page 78).

The rows show the ratio of full recovery, the average number of recovered directions, the average approximation error and the histograms (in log-scale) for the approximation errors in the case $R = 20, m_x = 20000$, from top to bottom respectively.
Figure 6.2.3: Distinguishing the exact directions based on the magnitude of the gradient \( \|\nabla f\|_2 \) along the exact directions for different degrees of quasi-orthogonality. Description on page 78
Figure 6.2.4: Distinguishing the approximations based on the magnitude of the gradient $\|\nabla f\|_2$ along the approximations for different degrees of quasi-orthogonality. Description on page 78.
Conclusion

The main purpose of this thesis was to see whether we can use the second derivative of a function

\[
 f(x) = \sum_{\ell=1}^{m_1} h_\ell(b_\ell^T \left( \begin{array}{c} g_1(a_1^T x) \\ g_2(a_2^T x) \\ \vdots \\ g_m(a_m^T x) \end{array} \right) )
\]

(6.2.1)

to identify the ridge directions \( a_1, \ldots, a_m, b_1, \ldots, b_m \). As we argued in the beginning, this type of function can describe a neural network with two hidden layers and one output. With this in mind, we want to conclude this thesis by describing the results with respect to neural networks, and focus less on the formal definition of the function \( f \). We can think of the ridge directions as the weights (weight vectors) of a neural network. It was shown that the Hessians of a neural network, represented by \( f \in \mathcal{F}_d(m_1, m) \) (cf. Definition 9), will to some degree concentrate around a space spanned by rank-1 matrices. By applying the theory developed in [4] to approximate these rank-1 basis elements, we get a lot of information about the weights of the neural network. In fact, if the approximations succeeds, we can identify the weights in the first layer up to a sign, whereas the identification of the weights in the second layer is only up to a transformation with a diagonal matrix \( A^T A = G_0 b_\ell \) (cf. chapter 2). In neural networks the functions \( g_i, h_\ell \) are usually given by an activation function \( \phi \) under slight shifts, the thresholds of the neurons, such that

\[
 G_0 = \left( \begin{array}{cccc} g'_1(0) & \cdots & \phi'(\theta_1) \\ \vdots & \ddots & \vdots \\ g'_m(0) & \cdots & \phi'(\theta_m) \end{array} \right).
\]

Hence, the identification of the weights in the second layer becomes possible once we know the thresholds \( \theta_1, \ldots, \theta_m \) of the neurons in the first layer. The theory in chapter 4 extends this results to the finite differences, and therefore, the whole approximation does not require the Hessians to be known. We assume that identifying \( G_0 \) becomes much easier if we restrict \( g_i, h_\ell \) to a family of smooth activation functions under slight shifts. Nevertheless, even without knowing \( G_0 \), our method still provides a lot of information about the weights of the network.
In the numerical experiments it became apparent that the quality of the results depends heavily on the domain from which the Hessians are sampled. For the type of network in our experiments, increasing the radius of the sphere resulted in a higher number of recovered direction and a lower approximation error. However, as we explained on page 76, the results illustrated within figure 6.1.6 indicate a certain trade-off: by increasing the radius of the sphere we got a better recovery of the elements $a_1, \ldots, a_m$ (corresponding to the weights of the first layer), but the recovery of $v_1, \ldots, v_{m_1}$ (corresponding to the weights of the second layer) got worse. This highlights the limitations of the uniform distribution on the sphere for our use case. Recovering the weights in an adaptive fashion could increase the consistency of our approach. Here, 'adaptive' means to change the distribution in Algorithm 1 or the initial values of Algorithm 2 based on the directions that are easy to recover.

One consequence of our procedure is that we need to distinguish to which layer the results of Algorithm 2 belong. In chapter 5 we proposed two methods: the first one was only based on the quasi-orthogonality of the vectors $a_1, \ldots, a_m, v_1, \ldots, v_{m_1}$, and the second one used the structure of a neural network with sigmoidal activations. This problematic is an important aspect with regards to future developments, namely the extension to more hidden layers. Assume we can identify a similar structure in the Hessians of deeper neural networks; i.e. we can associate the weights (represented by vectors) with rank-1 matrices and approximate their span by using finite differences. Then, by looking only at this matrix space, we can not decide which rank-1 matrix corresponds to a certain neuron (or layer). Therefore, this association has to be made based on their algebraic properties or the structure of the network. It is clear that this becomes increasingly difficult as we increase the number of layers.

Lastly, we want to share some thoughts on activation functions. In our examples we mostly considered sigmoidal activations, whereas many practitioners use piecewise linear activation functions; an example are networks with rectified linear units (ReLU), where the activation function is given by

$$x \mapsto \max(0, x).$$

We neglect for a moment that this function is not differentiable in 0. Naturally, these functions are problematic for methods that use higher order differentiation, since the second derivative is 0 on every linear segment. This is indirectly implied by our results in chapter 3. One of our assumptions (cf. Assumption 1) was that the $m + m_1$-th singular value of

$$E_{\mu_{g_1-1}} \left[ \text{vec}(H[f](X))^\otimes 2 \right] = \int_{g_{d-1}} \text{vec}(H[f](x))^\otimes 2 \, d\mu_{g_{d-1}}(x)$$

is well separated from 0, i.e. there is an $\alpha > 0$ such that

$$\sigma_{m+m_1} \left( E_{\mu_{g_1-1}} \left[ \text{vec}(H[f](X))^\otimes 2 \right] \right) \geq \alpha.$$
For functions of the type (6.2.2) the Hessians will almost be 0. Therefore, if there is an $\alpha$ such that (6.2.3) holds, it will be very small. As a consequence, our concentration bounds (e.g. Proposition 4) become very bad. For this reason, functions similar to (6.2.2) are not well suited for our approach.
Appendices
Appendix A

Definition A.1 (Beta distribution, (p.3)). The Beta(α, β) distribution for parameters α, β > 0 is characterized by a density on the segment [0, 1] given by:

\[
f(x) = \frac{1}{B(\alpha, \beta)} x^{\alpha-1}(1 - x)^{\beta-1},
\]

where \( B(\alpha, \beta) = \int_0^\infty x^{\alpha-1}(1 - x)^{\beta-1}dx = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha + \beta)} \). The mean of \( X \sim \text{Beta}(\alpha, \beta) \) is given by \( \mathbb{E}[X] = \frac{\alpha}{\alpha + \beta} \).

Theorem A.1 (General Hoeffding’s inequality,[13] Theorem 1.2.2). Let \( X_1, ..., X_m \) be independent, mean-zero, sub-gaussian random variables. Then, for every \( t \geq 0 \) we have

\[
P\left[ \left| \sum_{i=1}^m X_i \right| \geq t \right] \leq 2 \exp \left( -\frac{ct^2}{\sum_{i=1}^m \|X_i\|_{\psi_2}^2} \right).
\]

The following bound due to Wedin is a result on the stability of the singular value decomposition.

Theorem A.2 (Wedin’s bound[19]). Assume we have two real matrices \( A, \tilde{A} \) of the same size and with singular value decompositions

\[
A = (U_1 \quad U_2) \begin{pmatrix} \Sigma_1 & 0 \\ 0 & \Sigma_2 \end{pmatrix} \begin{pmatrix} V_1^T \\ V_2^T \end{pmatrix},
\]

\[
\tilde{A} = (\tilde{U}_1 \quad \tilde{U}_2) \begin{pmatrix} \hat{\Sigma}_1 & 0 \\ 0 & \hat{\Sigma}_2 \end{pmatrix} \begin{pmatrix} \hat{V}_1^T \\ \hat{V}_2^T \end{pmatrix},
\]

such that \( U_1, \tilde{U}_1 \) are of the same size. If there is an \( \bar{\alpha} > 0 \) such that

\[
\min_{k,k} |\sigma_k(\hat{\Sigma}_2) - \sigma_k(\Sigma_1)| \geq \bar{\alpha},
\]

and

\[
\min_k |\sigma_k(\Sigma_1)| \geq \bar{\alpha},
\]

87
then
\[ \|V_1V_1^T - \tilde{V}_1\tilde{V}_1^T\|_F \leq \frac{2}{\alpha} \|A - \tilde{A}\|_F. \]

**Theorem A.3.** [7, Theorem 5] All maximum independent sets of a bipartite graph can be generated in \( O(n^{2.5} + \gamma) \) where \( n \) is the number of vertices of the graph and \( \gamma \) the output size.

**Remark.** In the theorem above, the output size refers to the total sum of sizes of the output sets [7, p. 162]. The output sets are the maximum independent sets of the graph.

The following two results allow us to estimate the singular values of a sum of independent and semi-definite matrices. The first result was presented by Tropp [16] and works for the extreme singular values. The second one extends this result to the inner singular values and is due to Gittens and Tropp [5].

**Theorem A.4** (Matrix Chernov, [16, Corollary 5.2]). Let \( \{X_i| i = 1, \ldots, n\} \) be a sequence of independent, symmetric and positive semi-definite random matrices in \( \mathbb{R}^{d \times d} \) that satisfy
\[ \lambda_{\max}(X_i) \leq R \text{ almost surely.} \]

Let
\[ \mu_{\min} := \lambda_{\min} \left( \sum_i \mathbb{E}X_i \right), \]
and
\[ \mu_{\max} := \lambda_{\max} \left( \sum_i \mathbb{E}X_i \right), \]
where \( \lambda_{\min}, \lambda_{\max} \) denotes the minimal and maximal eigenvalue. Then
\[ \mathbb{P} \left( \lambda_{\min} \left( \sum_i X_i \right) \leq (1 - \delta)\mu_{\min} \right) \leq d \left[ \frac{e^{-\delta}}{(1 - \delta)^{1-\delta}} \right]^{\mu_{\min}/R} \text{ for } \delta \in [0, 1] \] and
\[ \mathbb{P} \left( \lambda_{\max} \left( \sum_i X_i \right) \geq (1 + \delta)\mu_{\max} \right) \leq d \left[ \frac{e^{\delta}}{(1 + \delta)^{1+\delta}} \right]^{\mu_{\max}/R} \text{ for } \delta \geq 0. \]

One standard simplification of the expressions in the last corollary is given as a remark in [16] cf. page 21:
\[ \mathbb{P} \left( \lambda_{\min} \left( \sum_i X_i \right) \leq (1 - \delta)\mu_{\min} \right) \leq de^{-(1-t)^2\mu_{\min}/2R} \text{ for } t \in [0, 1], \]
and
\[
\mathbb{P} \left( \lambda_{\text{max}} \left( \sum_{i} X_i \right) \geq (1 + \delta) \mu_{\text{max}} \right) \leq d \left[ \frac{e^{\gamma t \mu_{\text{max}} / R}}{t} \right] \quad \text{for } t \geq e.
\]

Before we state the version for the inner singular values, we have to give introduce some notation given in [5]. Let
\[
\mathbb{V}_d^n = \{ V \in \mathbb{C}^{n \times d} \mid V^* V = I \},
\]
where \( V^* \) is the hermitian transpose of \( V \). Next, define a function \( \Psi : \cup_{1 \leq k \leq n} \mathbb{V}_k^n \rightarrow \mathbb{R} \) such that
\[
\max_j \lambda_{\text{max}}(V^* X_j V) \leq \Psi(V) \quad \text{almost surely for each } V \in \cup_{1 \leq k \leq n} \mathbb{V}_k^n.
\]

Then the result is given as follows.

**Theorem A.5 ([5, Theorem 4.1]).** Consider a finite sequence \( \{ X_j \} \) of independent, random, positive-semidefinite matrices with dimension \( n \). Given any integer \( k \leq n \), define
\[
\mu_k = \lambda_k \left( \sum_j \mathbb{E}[X_j] \right),
\]
and let \( V_+ \in \mathbb{V}_{n-k+1}^n \) and \( V_- \in \mathbb{V}_k^n \) be isometric embeddings that satisfy
\[
\mu_k = \lambda_{\text{max}}(\sum_j V_+^* (\mathbb{E}[X_j]) V_+) = \lambda_{\text{min}}(\sum_j V_-^* (\mathbb{E}[X_j]) V_-).
\]

Then
\[
\mathbb{P} \left\{ \lambda_k(\sum_j X_j) \geq t \mu_k \right\} \leq (n - k + 1) \cdot \left[ \frac{e^{\gamma t \mu_k / \Psi(V_+)}}{t} \right] \quad \text{for } t \geq e, \quad \text{and}
\]
\[
\mathbb{P} \left\{ \lambda_k(\sum_j X_j) \leq t \mu_k \right\} \leq k \cdot e^{-(1-t)^2 \mu_k / (2 \Psi(V_-))},
\]

where \( \Psi \) is a function that satisfies \( \text{(A.0.2)} \).

As a simplification of the theory above, we can use a weaker estimate of \( \Psi(V_+), \Psi(V_-) \)

**Remark.** (cf. [5, Remark 4.1]) Note that
\[
\Psi(V_+) \leq \max_{V \in \mathbb{V}_{n-k+1}^n} \max_j \| V^* X_j V \| = \max_j \| X_j \|,
\]
\[
\Psi(V_-) \leq \max_{V \in \mathbb{V}_k^n} \max_j \| V^* X_j V \| = \max_j \| X_j \|.
\]
Theorem A.6 (Weyl’s inequality, [20]). Let $S, T$ be two symmetric $n \times n$ matrices, then for all $k = 1, \ldots, n$
\[
\lambda_k(S) + \lambda_n(T) \leq \lambda_k(S + T) \leq \lambda_k(S) + \lambda_1(T).
\] (A.0.3)
where $\lambda_k(S)$ denotes the $k$-th eigenvalue of $S$ and the eigenvalues of all involved matrices are sorted by their magnitude.

In most cases we will use the following corollary of Theorem A.6 which gives a similar result for the singular values.

Corollary A.1. Let $S, T$ be two symmetric $n \times n$ matrices, then for all $k = 1, \ldots, n$
\[
|\sigma_k(S) - \sigma_k(T)| \leq \|S - T\|,
\]
where $\sigma_k$ denotes the $k$-th singular value and $\|\cdot\|$ is the spectral norm.
Bibliography


