The condition number of Vandermonde matrices and its application to the stability analysis of a subspace method

Dissertation
zur Erlangung des Grades
Doktor der Naturwissenschaften (Dr. rer. nat.)
des Fachbereichs Mathematik/Informatik
der Universität Osnabrück

vorgelegt von

Dominik Nagel

Osnabrück, September 2020

Contents

Acknowledgments

First of all I would like to thank my whole family and particularly my siblings Tobias, Verena, Ramona, Julia, and Matthias and my parents Gabriele and Hermann for their constant and unconditioned support. My parents gave me the most valuable present a child can obtain from his or her parents besides being loved: the opportunity to develop freely and unconstrained. No less, I am exceedingly grateful for the love, patience and support from Julia. I appreciate the many time-outs and moments of recovering motivation experienced with all my friends, in particular, Toni, Nils, Hendrik, Patrick, Johannes, Maximilian and Sebastian.

Special thanks goes to my doctoral advisor Stefan Kunis for his valuable advice and support over the years. Uncountable discussions helped a lot finding and improving results and his encouraging especially after setbacks kept me going on with my research and writing this thesis. Anna Strotmann calculated a constant in her bachelor thesis, which I appreciate a lot, since her result leads to a nice bound that is included in this thesis. Furthermore, I gratefully acknowledge the willingness of Dmitry Batenkov to examine this thesis and I would like to thank Frank Filbir and André Uschmajew for inviting me to Munich and Leipzig, respectively, to present the actual state of my work.

I thank all people from the Institute of Mathematics at the Osnabrück University for the friendly and open atmosphere. Moreover, special thanks goes to all participants of the "coffee break" Markus, Stephan, Jonathan, Mathias, Alex G., Alex N., Arun, Carina, Jens, Daniel and Lorenzo for the joyful and relaxing time in- and outside university.

Finally, the funding by the graduate school DFG-GK1916 is gratefully appreciated.

Chapter 1

Introduction

In signal and image processing problems arise, in which sharp objects such as point sources have to be recovered. Frequently, the problems are motivated from physical applications like for instance astronomical imaging and microscopy in which point sources are stars or molecules. More abstractly, point sources can also represent components of signals like frequencies in sounds. In all applications, the lack of measurements or physical limitations obstructing the access of additional information lead to a limited resolution and make the task of localizing the point sources and determining their intensities extremely challenging. Two dominant resolution limits can be distinguished in this context. The first appears for instance in determining the exact position of a single small object in an image. This becomes extremely difficult if the image is blurred. Therefore, the localization of point sources with high precision from only partial information is called *weak super-resolution*. Secondly, for example the resolution limit in optical imaging of point sources are described by the classical Rayleigh and Abbe limit criteria [1, 77], see also the manuscript [24] and references therein. Due to the diffraction of the illuminating light, point sources can only be distinguished if they are at least separated by a distance proportional to the inverse frequency of light. Thus, the term *strong super-resolution* is used for resolving objects that are closer than such limit criteria impose. Both super-resolution tasks are covered by the term *sparse super-resolution* and by means of mathematical algorithms, it is possible to overcome resolution limits and efficiently recover signals that consist of point sources.

In sparse super-resolution problems, the signal is not described by a classical function. Since the signal is assumed to consists of point sources, it is typically modeled by a discrete complex Borel measure on the periodic unit interval

$$\mu := \sum_{j=1}^{M} \alpha_j \delta_{t_j},$$

where δ_{t_j} denotes the Dirac measure shifted to the point source location t_j and $\alpha_j \in \mathbb{C}$ is the respective coefficient or intensity. An example is given in Figure 1.0.1 (left). This model can be found in various fields and physical applications, e.g. spectral estimation, radar, sonar, direction of arrival estimation and super-resolution microscopy. We refer to [65, 25, 37] and references therein. The resolution limit caused by low quality devices and physical limitations, e.g. light diffraction in imaging, is modeled by a convolution of the measure with an appropriate function. Most simplified, this results in applying an ideal low-pass filter to the signal, see Figure 1.0.1 (middle), that cuts off high frequencies. Taking equispaced

samples of the signal, see Figure 1.0.1 (right), is therefore equivalent to observing its low-end frequencies. Transferred to the discrete measure μ, this means we obtain the data

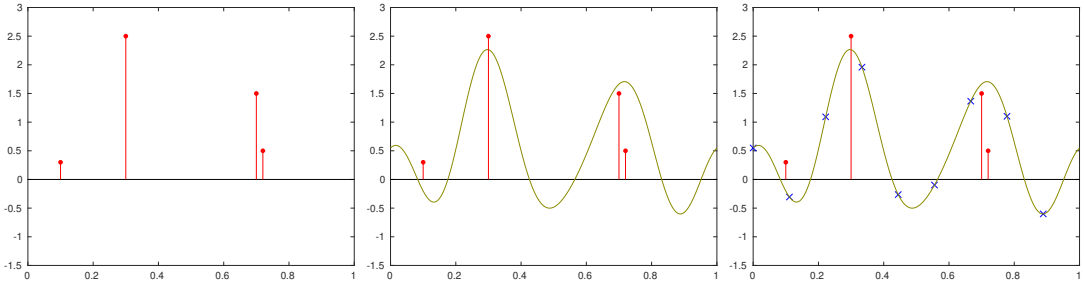


Figure 1.0.1: sketch of data acquisition in one dimensional sparse super-resolution; left: point sources; middle: low frequency version, observed "blurred image"; right: equispaced samples, equivalent to low frequencies.

$$\widehat{\mu}(k) := \int_{\mathbb{T}} \mathrm{e}^{-2\pi i k t} \, d\mu(t) = \sum_{j=1}^{M} \alpha_j \mathrm{e}^{-2\pi i k t_j}, \quad \text{for} \quad k = 0, \dots, N-1. \tag{1.0.1}$$

The advantage of this model is that point sources are not assumed to lie on a predefined grid making therefore weak super-resolution possible. Moreover, from mathematical perspective, the exact data from (1.0.1) in combination with the assumption that the signal can be modeled by μ even enables strong super-resolution for arbitrary small separation distances and with infinite precision. Although there is a non-linear relationship in (1.0.1) between the unknowns α_j, t_j and the data $\widehat{\mu}(k)$, already in the 18th century de Prony came up with a method in [30] that allows to exactly solve this problem (nowadays known as Prony's method) with only $N = 2M$ samples, assuming that the number of parameters M is known a priori. Prony's method consists of two steps. The first is to reconstruct the point source locations via computing the *nodes* $z_j := \mathrm{e}^{2\pi i t_j}$ that appear as roots of a certain polynomial and in the second step the coefficients are calculated. Since the nodes are known after the first step, the latter is done by solving the (over-determined) linear system of equations given by (1.0.1) in which the system matrix is given by

$$\boldsymbol{A} = \left(z_j^k\right)_{\substack{j=1,\dots,M \\ k=0,\dots,N-1}},$$

a rectangular Vandermonde matrix of degree $N-1$ with nodes on the complex unit circle. Such matrices are central objects of this thesis.

So far the data in (1.0.1) was assumed to be exact, but in applications there are always errors when measuring or storing data. For this situation the development of more robust algorithms started in the second half of the 20th century. Besides optimization approaches [23, 83, 32, 33], subspace methods were invented that reconstruct the nodes via solving eigenvalue or generalized eigenvalue problems, such as the matrix pencil (MP) method [56], the multiple signal classification (MUSIC) algorithm [99] and the estimation of signal parameters via rotational invariance techniques (ESPRIT) method [96], which are closely related to Prony's method. We also refer to [90, 91, 105] and references therein. The question then arises, how capable these algorithms are in the different super-resolution scenarios.

All methods have in common that they roughly perform in the two steps described above. The stability analysis of solving the Vandermonde system for reconstructing the coefficients, utilizing a standard least squares method, strongly relies on the condition number of the involved Vandermonde matrix. Moreover, recently a stronger focus was put on the stability analysis of the subspace methods in which errors in the data (1.0.1) are only assumed to be bounded. Results from [80, 6, 5, 74] show that the error amplification when reconstructing the location of point sources can be controlled by an expression involving the condition number of such Vandermonde matrices as well.

Therefore, studying the condition number of Vandermonde matrices with nodes on the unit circle is of central importance. In the univariate case, they have been studied intensively if the nodes are well-separated, i.e., the minimal separation distance of the nodes is greater than the inverse degree of the Vandermonde matrix, see [80, 7, 34]. In this scenario, the Vandermonde matrices are conditioned well leading to mild error amplifications using subspace methods. This shows that these methods have excellent weak super-resolution abilities. On the contrary, the situation where the separation is smaller than the inverted matrix degree and multivariate generalizations are topic of recent and ongoing research [12, 11, 73, 69, 68]. This is where most contributions of this thesis lie. Furthermore, these results enable to analyze strong super-resolution capabilities of subspace methods as already started in [13, 72, 74].

Contributions

The main contributions of this thesis are the following:

i) In Theorem 3.4.12 we prove a lower bound on the smallest singular value for multivariate Vandermonde matrices (having a certain max-degree) with nodes component-wise on the complex unit circle that build several clusters. Therefore, we extend univariate results from [11, 12, 73]. In contrast to [73], where our applied technique is developed for the univariate case, assumptions on the cluster separation and the lower bound on the singular value are not dependent on the total number of nodes. Instead the number of nodes in the largest cluster comes into play. Moreover, our result not only generalizes the univariate case but also improves upon it. This improvement is mainly based on a careful analysis of powers of modified Dirichlet kernels in Lemma 2.4.11 and the application of their decay properties and packing arguments to the column-sum norm of certain matrices. These results are published in [69]. A new upper bound on the largest singular value of such matrices (Lemma 3.4.6), also dependent on the number of nodes in the largest cluster instead of the total number of nodes, yields, together with the above, an upper bound for the condition number with same dependencies.

ii) For multivariate Vandermonde matrices with well-separated nodes, we provide lower and upper bounds for the extremal singular values in Section 3.5. In particular, for the first time quantitative lower bounds for the smallest singular value are proven for the multivariate case. This extends univariate results from [15, 41, 76, 7, 34].

iii) In the univariate case, we investigate the condition number of Vandermonde matrices with several pair clusters, i.e., clusters consist of at most two nodes, in Section 3.3. A Schur-decomposition technique allows to divide the node set into two sets of well-separated nodes and to apply known bounds for corresponding Vandermonde matrices

in order to find bounds for the largest and smallest singular values. We additionally investigate the limitation of this technique when applying it to larger clusters or higher dimensions. These results are published in [68].

iv) In contrast to statistical and asymptotic stability analysis the deterministic stability analysis (assuming the noise is bounded) of subspace methods made great progress in the last years. For the matrix pencil method and a modified version of it, first results were presented in [6, 5] and [80]. The deterministic stability of the ESPRIT method was investigated in [16, 94, 6, 5, 74]. Particularly, in the last three references the error amplification is given in terms of the extremal singular values of involved Vandermonde matrices. The two approaches are based on the same tools from matrix perturbation theory but while [7] identifies a matrix valued least squares problem, [74] uses principal vectors and angles of subspaces to refine bounds, so that both results in their stated form seem to be hardly to compare. In Chapter 4, we present the stability analysis of the ESPRIT algorithm mainly following [74] in an improved way (most importantly we are able to drop a factor which is the square root of total number of nodes). Afterwards, we show (see Remark 4.3.6) that until a certain point the approach from [74] provides basically the same result as in [6]. Furthermore, we follow [94] and present a stability analysis for reconstructing coefficients.

Outline of the thesis

Chapter 2 contains preliminaries and notations for making the thesis containing most and frequently used arguments. We start with basic linear algebra tools, vector and matrix analysis, followed by more advanced matrix perturbation theory. We additionally focus on the detailed description of principal angles and principal vectors between given subspaces of finite dimensional vector spaces. Especially, the careful study of matrices with columns being principal vectors allows to slightly improve results on the stability analysis of the ESPRIT algorithm in Chapter 4 compared to the one from [74]. A summary of necessary Fourier analysis is given afterwards and finally auxiliary functions and their properties are stated. In particular analytic properties of the Dirichlet kernel, its derivatives and powers of multivariate modified Dirichlet kernels are elaborated. These technical results may be of interest separately.

Chapter 3 deals with the condition number of Vandermonde matrices. Firstly, we survey the research on square Vandermonde matrices with real and complex nodes. There we see that Vandermonde matrices with real nodes have exponentially growing condition number with respect to the matrix size. In contrast, if nodes are on the complex unit circle, the situation may be better, even perfectly conditioned Vandermonde matrices exist. The remaining chapter deals with rectangular Vandermonde matrices having more columns than rows (the latter correspond to the different nodes). A short recapitulation of recent results for nodes on the complex unit disc is given. Afterwards, we restrict to the situation where nodes are from the complex unit circle which does not change throughout the thesis from then on. If nodes are well-separated with respect to the inverse of the Vandermonde matrix degree the condition number of that matrix is small. The situation where nodes are not well-separated anymore was investigated during the last years and is still ongoing research. Clustered node configurations turned out to be a reasonable model for describing such node sets and being

beneficial for estimating condition numbers of the corresponding Vandermonde matrices. We start with the situation of one cluster with two nodes that is among well-separated nodes and then extent it to node sets consisting of several pair clusters. Lower and upper bounds for the extremal singular values and hence for the condition number of corresponding Vandermonde matrices are given. Limitations of the technique for multivariate extensions and its extension to node sets with clusters consisting of more than two nodes are also discussed. Afterwards, we switch over to the multivariate setting. We present a multivariate extension of the technique developed in [73] that allows to bound the condition number of multivariate Vandermonde matrices with clustered node configurations. Results for the multivariate setting are completely new as far as we know. In the univariate case in general and for pair clusters in arbitrary dimensions, new and known upper bounds for the smallest singular value complete the picture how the smallest singular value and hence the condition number behaves. Furthermore, an example shows that for larger clusters in higher dimensions the geometric configuration within the cluster plays an additional role for the dependency of the smallest singular value on the node separation and the Vandermonde matrix degree. We recap existing results for univariate clustered node configurations and compare them with our own results. Finally, known results for the case of well-separated nodes in multiple dimensions are collected, refined and extended.

Chapter 4 serves mainly for applying the condition number bounds from Chapter 3. We present a stability analysis of the univariate ESPRIT method based on ideas from [6, 5] and oriented at [74] with some modifications and improvements. First of all, we recapitulate the method in context of reconstructing exponential sums. Afterwards, the stability analysis is given for the ESPRIT method that reconstructs the nodes of an exponential sum. Additionally, we analyze the stability of the coefficient reconstruction following [94].

All images and sketches in the thesis except for the university logo on the title page are produced with MATLAB, Inkscape and TikZ.

Chapter 2

Preliminaries and Notation

In this chapter, we introduce and present most of the tools we need throughout the thesis. Basic mathematical notions are additionally collected in order to fix notation. Simple proofs are included for the convenience of the reader. Longer proofs are omitted, but appropriate references are given. We start with vector and matrix analysis followed by matrix perturbation results. Afterwards Fourier analytic tools are recapitulated and finally, auxiliary functions, especially trigonometric kernel functions and their analytic properties are addressed. If the reader is familiar with these topics, it is possible to treat this chapter as mathematical tool box which is referred to in the course of the thesis. For this purpose most important notations can also be found in the glossary at the end.

2.1 Vector and Matrix analysis

We start this section by introducing some notations and basic results concerning vectors and matrices. Then we recap some results for eigenvalues of matrices and in particular for Hermitian matrices. Afterwards, the singular value decomposition, connections of singular values to eigenvalues and some auxiliary results in connection to norms are given. In the end the Moore-Penrose pseudo inverse and the definition of the condition number of matrices are stated.

Vectors and Norms

An appropriate reference for this section is [103]. Throughout the thesis, we often deal with vectors from the finite dimensional complex vector space $\mathbb{C}^m$, for some $m \in \mathbb{N}$, and denote them by boldface lower-case letters, e.g. $\boldsymbol{v} \in \mathbb{C}^m$. They are regarded as column vectors. The conjugate vector is denoted by $\bar{\boldsymbol{v}}$, the transpose by $\boldsymbol{v}^\top$ and the conjugate transpose by $\boldsymbol{v}^*$. We access the components by using brackets with the respective index, e.g. $(\boldsymbol{v})_j$ for the j-th component of $\boldsymbol{v}$, unless the components are defined explicitly, for instance by $\boldsymbol{v} = (v_1, \dots, v_m)^\top$. The vector spaces are known to be Hilbert spaces with respect to the Euclidean scalar product

$$\langle \cdot, \cdot \rangle : \mathbb{C}^m \times \mathbb{C}^m \to \mathbb{C}, \quad \langle \boldsymbol{u}, \boldsymbol{v} \rangle := \sum_{k=1}^{m} (\boldsymbol{u})_k \, \overline{(\boldsymbol{v})_k} = \boldsymbol{v}^* \boldsymbol{u}, \tag{2.1.1}$$

where the most right hand side is due to the standard row vector times column vector multiplication. This is a helpful and compact way of writing the scalar product later on. Adding one of the well-known norms, that are functions of the form $\mathbb{C}^m \to \mathbb{R}_{\geq 0}$, the $\mathbb{C}^m$ becomes a normed vector space. In this thesis the following norms appear.

Definition 2.1.1 (Vector norms)**.**
For $m \in \mathbb{N}_+$ consider the complex normed vector space $\mathbb{C}^m$ and let $\boldsymbol{v} \in \mathbb{C}^m$. Then we have the norms

$$\|\boldsymbol{v}\| := \left(\sum_{k=1}^{m} |(\boldsymbol{v})_k|^2\right)^{\frac{1}{2}}, \quad \|\boldsymbol{v}\|_1 := \sum_{k=1}^{m} |(\boldsymbol{v})_k|, \quad \text{and} \quad \|\boldsymbol{v}\|_\infty := \max_{1\leq k\leq m} |(\boldsymbol{v})_k|.$$

The first norm is the Euclidean or 2-norm, the second is called 1-norm and third is the max- or ∞-norm.

Since the vector space $\mathbb{C}^m$ is finite dimensional, all vector norms are equivalent. This means that each norm is larger than a constant times another norm. The norms from Definition 2.1.1 have the following relations.

Lemma 2.1.2 (Norm equivalences, cf. [49, Sec. 2.2.2])**.**
Let $m \in \mathbb{N}$ and $\boldsymbol{u}, \boldsymbol{v} \in \mathbb{C}^m$ be arbitrary. Then we have

$$\|\boldsymbol{v}\|_\infty \leq \|\boldsymbol{v}\| \leq \|\boldsymbol{v}\|_1 \leq \sqrt{m}\,\|\boldsymbol{v}\| \leq m\,\|\boldsymbol{v}\|_\infty.$$

A useful tool that finds place in many proofs of inequalities is the Cauchy-Schwarz inequality.

Lemma 2.1.3 (Cauchy-Schwarz inequality for vectors, [102, Ch. 1])**.**
Let $m \in \mathbb{N}$ and $\boldsymbol{u}, \boldsymbol{v} \in \mathbb{C}^m$. Then the Cauchy-Schwarz inequality

$$|\langle \boldsymbol{u}, \boldsymbol{v}\rangle| \leq \|\boldsymbol{u}\|\,\|\boldsymbol{v}\|$$

holds.

Matrices

Since notation related to matrix analysis varies a lot, we start by fixing it for basic notions. All these concepts can be found in detail for example in [49, 55]. Let $m, n \in \mathbb{N}$. Similar to vectors, we denote the vector space of complex matrices by $\mathbb{C}^{m\times n}$ and individual matrices by boldface capital letters like $\boldsymbol{M} \in \mathbb{C}^{m\times n}$. Again, unless not defined differently, we use brackets, but this time with two indices, to access the elements of a matrix. For instance, let $j, k \in \mathbb{N}$ with $1 \leq j \leq m$ and $1 \leq k \leq n$, then the element in the j-th row and k-th column of $\boldsymbol{M}$ is given by $(\boldsymbol{M})_{j,k}$. In Section 2.1, vectors in $\mathbb{C}^m$ are regarded as column vectors, now they can be considered equivalently as matrices in $\mathbb{C}^{m\times 1}$. Therefore, we use the common matrix multiplication, a map of the form $\mathbb{C}^{m\times n} \times \mathbb{C}^{n\times k} \to \mathbb{C}^{m\times k}$, and the usual notation for multiplication (omitting the dot most of the times). In particular, this defines multiplication of two vectors $\boldsymbol{u} \in \mathbb{C}^m, \boldsymbol{v} \in \mathbb{C}^n$ of the form $\boldsymbol{u}\boldsymbol{v}^*$ (often called outer product), which results in a matrix in $\mathbb{C}^{m\times n}$. One has to note that this is therefore not a commutative operation in general. The complex conjugation operator is used component-wise, i.e. for

$\boldsymbol{M} = \left(M_{j,k}\right)_{j,k=1}^{m,n} \in \mathbb{C}^{m\times n}$, we define $\overline{\boldsymbol{M}} := \left(\overline{M_{j,k}}\right)_{j,k=1}^{m,n}$. The rank of a matrix is given by $\operatorname{rank}(\boldsymbol{M})$ and in the case that a quadratic matrix $\boldsymbol{M} \in \mathbb{C}^{m\times m}$ is invertible we call it regular and its inverse is denoted by $\boldsymbol{M}^{-1}$. We recall that a quadratic matrix $\boldsymbol{M} \in \mathbb{C}^{m\times m}$ is called *unitary* if and only if $\boldsymbol{M}^* = \boldsymbol{M}^{-1}$. The eigenvalues of a quadratic matrix are denoted by $\lambda_j(\boldsymbol{M}), j = 1, \ldots, m$, unless given differently.

If $\boldsymbol{M}$ is *Hermitian*, i.e. $\boldsymbol{M} = \boldsymbol{M}^*$, the eigenvalues are all real and we use $\lambda_{\max}(\boldsymbol{M})$ and $\lambda_{\min}(\boldsymbol{M})$ for the largest and smallest eigenvalue, respectively. We call a Hermitian matrix positive (semi-)definite if all its eigenvalues are greater (or equal) zero. The operator $\operatorname{diag}(a_1, \ldots, a_m)$ produces quadratic, diagonal matrix in $\mathbb{C}^{m\times m}$ with diagonal entries $a_1, \ldots, a_m$. If the matrix is specified as an element of a different matrix space, e.g. in $\mathbb{C}^{m\times n}$, then it puts the entries on the diagonal starting in the upper left corner. The remaining parts of the matrix are filled with zeros. The range of a matrix is given by $\operatorname{range}(\boldsymbol{M})$ and equals the linear span of its columns. Finally, the kernel of a matrix is denoted by $\ker(\boldsymbol{M})$.

As a tool for later proofs, we state the famous Courant-Fischer min-max theorem that provides a variational characterization for eigenvalues of Hermitian matrices. The proof can be found in the given reference.

Theorem 2.1.4 (Courant-Fischer, cf. [55, Thm. 4.2.6])**.**
Let $\boldsymbol{M} \in \mathbb{C}^{m\times m}$ be Hermitian and let $\lambda_1 \geq \cdots \geq \lambda_m$ be its eigenvalues. Let $\mathcal{S}$ denote a subspace of $\mathbb{C}^m$. Then for each $k = 1, \ldots, m$ we have

$$\lambda_k = \min_{[\mathcal{S}\colon \dim \mathcal{S} = m-k+1]} \max_{[\boldsymbol{x}\in\mathcal{S}, \boldsymbol{x}\neq\boldsymbol{0}]} \frac{\boldsymbol{x}^*\boldsymbol{M}\boldsymbol{x}}{\boldsymbol{x}^*\boldsymbol{x}} = \min_{[\mathcal{S}\colon \dim \mathcal{S} = m-k+1]} \max_{[\boldsymbol{x}\in\mathcal{S}, \|\boldsymbol{x}\|=1]} \boldsymbol{x}^*\boldsymbol{M}\boldsymbol{x}$$

and

$$\lambda_k = \max_{[\mathcal{S}\colon \dim \mathcal{S} = k]} \min_{[\boldsymbol{x}\in\mathcal{S}, \boldsymbol{x}\neq\boldsymbol{0}]} \frac{\boldsymbol{x}^*\boldsymbol{M}\boldsymbol{x}}{\boldsymbol{x}^*\boldsymbol{x}} = \max_{[\mathcal{S}\colon \dim \mathcal{S} = k]} \min_{[\boldsymbol{x}\in\mathcal{S}, \|\boldsymbol{x}\|=1]} \boldsymbol{x}^*\boldsymbol{M}\boldsymbol{x}$$

In particular, we have the Rayleigh-Ritz characterizations of the extremal eigenvalues

$$\lambda_{\max}(\boldsymbol{M}) = \max_{\boldsymbol{x}\in\mathbb{C}^m, \boldsymbol{x}\neq 0} \frac{\boldsymbol{x}^*\boldsymbol{M}\boldsymbol{x}}{\boldsymbol{x}^*\boldsymbol{x}}, \quad \textit{and} \quad \lambda_{\min}(\boldsymbol{M}) = \min_{\boldsymbol{x}\in\mathbb{C}^m, \boldsymbol{x}\neq 0} \frac{\boldsymbol{x}^*\boldsymbol{M}\boldsymbol{x}}{\boldsymbol{x}^*\boldsymbol{x}}.$$

Additionally, we state the following theorem and a lemma in connection with eigenvalues of Hermitian matrices.

Theorem 2.1.5 (Weyl inequalities , cf. [55, Thm. 4.3.1])**.**
Let $\boldsymbol{M}, \boldsymbol{N} \in \mathbb{C}^{m\times m}$ be Hermitian matrices and let $\lambda_1(\boldsymbol{M}) \geq \cdots \geq \lambda_m(\boldsymbol{M})$ and $\lambda_1(\boldsymbol{N}) \geq \cdots \geq \lambda_m(\boldsymbol{N})$ be their eigenvalues. Then for each $j = 1, \ldots, m$ we have

$$\lambda_{j-\ell+1}(\boldsymbol{M}) + \lambda_\ell(\boldsymbol{N}) \geq \lambda_j(\boldsymbol{M}+\boldsymbol{N}) \geq \lambda_{j+k}(\boldsymbol{M}) + \lambda_{m-k}(\boldsymbol{N}),$$

where $\ell = 1, \ldots, j$ and $k = 0, \ldots, m-j$.

Lemma 2.1.6 (Inclusion principle, cf. [55, Thm. (4.3.28)])**.**
Let $\boldsymbol{M} \in \mathbb{C}^{m\times m}$ be Hermitian with block partition

$$\boldsymbol{M} = \begin{pmatrix} \boldsymbol{M}_1 & \boldsymbol{M}_2 \\ \boldsymbol{M}_2^* & \boldsymbol{M}_3 \end{pmatrix}, \quad \boldsymbol{M}_1 \in \mathbb{C}^{n\times n}, \boldsymbol{M}_2 \in \mathbb{C}^{n\times(m-n)}, \boldsymbol{M}_3 \in \mathbb{C}^{(m-n)\times(m-n)}.$$

Let $\lambda_1(\boldsymbol{M}) \geq \cdots \geq \lambda_m(\boldsymbol{M})$ and $\lambda_1(\boldsymbol{M}_1) \geq \cdots \geq \lambda_m(\boldsymbol{M}_1)$ be the eigenvalues of $\boldsymbol{M}$ and $\boldsymbol{M}_1$. Then for $j = 1, \ldots, n$ we have

$$\lambda_{j+m-n}(\boldsymbol{M}) \geq \lambda_j(\boldsymbol{M}_1) \geq \lambda_j(\boldsymbol{M}).$$

In particular, we have

$$\lambda_{\max}(\boldsymbol{M}) \geq \lambda_{\max}(\boldsymbol{M}_1) \geq \lambda_{\min}(\boldsymbol{M}_1) \geq \lambda_{\min}(\boldsymbol{M}).$$

The space of matrices together with an appropriate norm is a normed vector space. Most of the time, we make use of the spectral norm. Moreover, the row- and column-sum norms, also known as ∞- and 1-norms appear as well. These three norms are induced by the vector norms from Definition 2.1.1. Additionally, the Frobenius norm is used at some points. Notation for the matrix norms is introduced in the next definition.

Definition 2.1.7 (Matrix norms, cf. [55, Sec. 5.6], [49, Sec. 2.3.2])**.**
Let $m, n \in \mathbb{N}$ and $\boldsymbol{M} \in \mathbb{C}^{m\times n}$. Then we denote

i) the spectral norm *of $\boldsymbol{M}$ by*

$$\|\boldsymbol{M}\| := \max_{\boldsymbol{x}\in\mathbb{C}^n} \frac{\|\boldsymbol{M}\boldsymbol{x}\|}{\|\boldsymbol{x}\|} = \max_{\boldsymbol{x}\in\mathbb{C}^n, \|\boldsymbol{x}\|=1} \|\boldsymbol{M}\boldsymbol{x}\| = \sqrt{\lambda_{\max}(\boldsymbol{M}^*\boldsymbol{M})},$$

ii) the ∞-norm *or* row-sum norm *by*

$$\|\boldsymbol{M}\|_\infty := \max_{\boldsymbol{x}\in\mathbb{C}^n} \frac{\|\boldsymbol{M}\boldsymbol{x}\|_\infty}{\|\boldsymbol{x}\|_\infty} = \max_{\boldsymbol{x}\in\mathbb{C}^n, \|\boldsymbol{x}\|_\infty=1} \|\boldsymbol{M}\boldsymbol{x}\|_\infty = \max_{1\leq j\leq m} \sum_{k=1}^{n} \left|(\boldsymbol{M})_{j,k}\right|,$$

iii) the 1-norm *or* column-sum norm *by*

$$\|\boldsymbol{M}\|_1 := \max_{\boldsymbol{x}\in\mathbb{C}^n} \frac{\|\boldsymbol{M}\boldsymbol{x}\|_1}{\|\boldsymbol{x}\|_1} = \max_{\boldsymbol{x}\in\mathbb{C}^n, \|\boldsymbol{x}\|_1=1} \|\boldsymbol{M}\boldsymbol{x}\|_1 = \max_{1\leq k\leq n} \sum_{j=1}^{m} \left|(\boldsymbol{M})_{j,k}\right|.$$

iv) and the Frobenius norm *by*

$$\|\boldsymbol{M}\|_{\mathrm{F}} := \left(\sum_{j=1}^{m}\sum_{k=1}^{n} \left|(\boldsymbol{M})_{j,k}\right|^2\right)^{\frac{1}{2}} = \left(\sum_{j=1}^{n} \lambda_j(\boldsymbol{M}^*\boldsymbol{M})\right)^{\frac{1}{2}}.$$

The Frobenius norm is basically the Euclidean norm of the vector containing all entries of the matrix $\boldsymbol{M}$. The spectral norm and the Frobenius norm are unitary invariant, which follows directly from the respective last identities in their definitions. When multiplying $\boldsymbol{M}$ with a unitary matrix from right we use that eigenvalues are invariant under similarity transformation and when multiplying from left we use the characterizing property of unitary matrices. The first three norms are induced by vector norms and hence, they are consistent with the respective vector norms, i.e. $\|\boldsymbol{M}\boldsymbol{v}\| \leq \|\boldsymbol{M}\|\,\|\boldsymbol{v}\|$, $\|\boldsymbol{M}\boldsymbol{v}\|_\infty \leq \|\boldsymbol{M}\|_\infty\,\|\boldsymbol{v}\|_\infty$ and $\|\boldsymbol{M}\boldsymbol{v}\|_1 \leq \|\boldsymbol{M}\|_1\,\|\boldsymbol{v}\|_1$. Similar to vector norms matrix norms are equivalent and the relations between the above defined matrix norms are given next.

Lemma 2.1.8 (Matrix norm relations, cf. [49, Sec. 2.3.2, 2.3.3])**.**
Let $m, n \in \mathbb{N}$ and $\boldsymbol{M} \in \mathbb{C}^{m\times n}$ a matrix with $\operatorname{rank}(\boldsymbol{M}) = r$*. Then we have*

$$\|\boldsymbol{M}\| \leq \|\boldsymbol{M}\|_{\mathrm{F}} \leq \sqrt{r}\,\|\boldsymbol{M}\|,$$

$$\frac{1}{\sqrt{n}} \|\boldsymbol{M}\|_\infty \leq \|\boldsymbol{M}\| \leq \sqrt{m} \|\boldsymbol{M}\|_\infty ,$$
$$\frac{1}{\sqrt{m}} \|\boldsymbol{M}\|_1 \leq \|\boldsymbol{M}\| \leq \sqrt{n} \|\boldsymbol{M}\|_1 ,$$

and

$$\|\boldsymbol{M}\| \leq \sqrt{\|\boldsymbol{M}\|_1 \|\boldsymbol{M}\|_\infty}.$$

Singular values

Next, we look at the singular value decomposition of a matrix and variants of that. Such a decomposition exists for every matrix in $\mathbb{C}^{m\times n}$ and provides a powerful tool for analysis of matrices and related objects.

Theorem 2.1.9 (Singular value decomposition, [104, Ch. 1, Thm. 4.1], [49, Thm. 2.5.2])**.** *Let $\boldsymbol{M} \in \mathbb{C}^{m\times n}$ a matrix with* $\operatorname{rank}(\boldsymbol{M}) = r$ *and let $s = \min\{m, n\}$. Then there exist unitary matrices $\boldsymbol{U}_0 \in \mathbb{C}^{m\times m}$, $\boldsymbol{V}_0 \in \mathbb{C}^{n\times n}$ and a diagonal matrix $\boldsymbol{\Sigma}_0 := \operatorname{diag}(\sigma_1, \ldots, \sigma_s) \in \mathbb{C}^{m\times n}$ with singular values $\sigma_1 \geq \cdots \geq \sigma_r > 0 = \sigma_{r+1} = \cdots = \sigma_s \in \mathbb{R}_{\geq 0}$, such that we can decompose $\boldsymbol{M}$ to*

$$\boldsymbol{M} = \boldsymbol{U}_0 \boldsymbol{\Sigma}_0 \boldsymbol{V}_0^* = \boldsymbol{U}_0 \begin{pmatrix} \sigma_1 & & & \\ & \ddots & & \boldsymbol{0}_{r\times(n-r)} \\ & & \sigma_r & \\ \boldsymbol{0}_{(m-r)\times r} & & & \boldsymbol{0}_{(m-r)\times(n-r)} \end{pmatrix} \boldsymbol{V}_0^*. \tag{2.1.2}$$

This is called the singular value decomposition (SVD) *of the matrix $\boldsymbol{M}$. For the j-th singular value of a matrix $\boldsymbol{M}$, we write $\sigma_j(\boldsymbol{M})$ unless specified differently. We denote the largest singular value by $\sigma_{\max}(\boldsymbol{M})$ and the smallest singular value by $\sigma_{\min}(\boldsymbol{M})$. The columns of $\boldsymbol{U}_0$ and $\boldsymbol{V}_0$ are called* left *and* right singular vectors*, respectively.*

Proof. There are several ways to proof the existence of a singular value decomposition in the literature. The proof in [104, Ch. 1, Thm. 4.1] is written down for complex matrices and uses the matrices $\boldsymbol{M}^*\boldsymbol{M}$ and $\boldsymbol{M}\boldsymbol{M}^*$. They are Hermitian and positive definite and hence, have non-negative eigenvalues and unitary eigenvalue decompositions. Furthermore, the eigenvalues are, except for additional zeros, the same.

Here, we follow the proof given in [49, Thm. 2.5.2]. It uses a constructive approach and is written down for the real case, but the extension to the complex case is analogous. If $\boldsymbol{M} = \boldsymbol{0}_{m\times n}$ nothing is to show. So let $\boldsymbol{M} \neq \boldsymbol{0}_{m\times n}$, $\boldsymbol{u}_1 \in \mathbb{C}^m$ and $\boldsymbol{v}_1 \in \mathbb{C}^n$ be vectors with $\|\boldsymbol{u}_1\| = \|\boldsymbol{v}_1\| = 1$ such that $\boldsymbol{M}\boldsymbol{v}_1 = \sigma_1\boldsymbol{u}_1$, where $\sigma_1 := \|\boldsymbol{M}\| \neq 0$. They always exist by definition of the spectral norm. Since a unit norm vector can be extended to an orthonormal basis for its space, we can find matrices $\widetilde{\boldsymbol{U}}_1 \in \mathbb{C}^{m\times(m-1)}$ and $\widetilde{\boldsymbol{V}}_1 \in \mathbb{C}^{n\times(n-1)}$ such that $\boldsymbol{U}_1 = \begin{pmatrix} \boldsymbol{u}_1 & \widetilde{\boldsymbol{U}}_1 \end{pmatrix} \in \mathbb{C}^{m\times m}$ and $\boldsymbol{V}_1 = \begin{pmatrix} \boldsymbol{v}_1 & \widetilde{\boldsymbol{V}}_1 \end{pmatrix} \in \mathbb{C}^{n\times n}$ are unitary. By the choice of $\boldsymbol{u}_1$ and $\boldsymbol{v}_1$, we have

$$\boldsymbol{U}_1^*\boldsymbol{M}\boldsymbol{V}_1 = \begin{pmatrix} \boldsymbol{u}_1^* \\ \widetilde{\boldsymbol{U}}_1^* \end{pmatrix} \boldsymbol{M} \begin{pmatrix} \boldsymbol{v}_1 & \widetilde{\boldsymbol{V}}_1 \end{pmatrix} = \begin{pmatrix} \boldsymbol{u}_1^* \\ \widetilde{\boldsymbol{U}}_1^* \end{pmatrix} \begin{pmatrix} \sigma_1\boldsymbol{u}_1 & \boldsymbol{M}\widetilde{\boldsymbol{V}}_1 \end{pmatrix} = \begin{pmatrix} \sigma_1 & \boldsymbol{w}^* \\ \boldsymbol{0}_{(m-1)\times 1} & \boldsymbol{M}_1 \end{pmatrix}, \tag{2.1.3}$$

where $\boldsymbol{w} \in \mathbb{C}^{n-1}$ and $\boldsymbol{M}_1 \in \mathbb{C}^{(m-1)\times(n-1)}$. Furthermore, since the spectral norm is unitary invariant, using its definition leads to

$$\sigma_1 = \|\boldsymbol{M}\| = \left\| \begin{pmatrix} \sigma_1 & \boldsymbol{w}^* \\ \boldsymbol{0}_{(m-1)\times 1} & \boldsymbol{M}_1 \end{pmatrix} \right\| \geq \left\| \begin{pmatrix} \sigma_1 & \boldsymbol{w}^* \\ \boldsymbol{0}_{(m-1)\times 1} & \boldsymbol{M}_1 \end{pmatrix} \begin{pmatrix} \sigma_1 \\ \boldsymbol{w} \end{pmatrix} \right\|$$

$$= \left\| \begin{pmatrix} \sigma_1^2 + \boldsymbol{w}^*\boldsymbol{w} \\ \boldsymbol{M}_1\boldsymbol{w} \end{pmatrix} \right\| \geq \sqrt{\sigma_1^2 + \boldsymbol{w}^*\boldsymbol{w}}$$

and therefore, necessarily $\boldsymbol{w} = \boldsymbol{0}$ in (2.1.3). If $n = 1$ or $m = 1$ the proof is finished. Otherwise if $n, m > 1$, we can proceed by induction. Assume we have unitary matrices $\boldsymbol{U}_2 \in \mathbb{C}^{(m-1)\times m-1}, \boldsymbol{V}_2 \in \mathbb{C}^{(n-1)\times(n-1)}$ such that $\boldsymbol{U}_2^*\boldsymbol{M}_1\boldsymbol{V}_2 = \boldsymbol{\Sigma}_2$, where $\boldsymbol{\Sigma}_2 \in \mathbb{C}^{(m-1)\times(n-1)}$ is diagonal with entries in $\mathbb{R}_{\geq 0}$. For the largest entry σ_2 of $\boldsymbol{\Sigma}_2$ it holds by using (2.1.3)

$$\begin{aligned} \sigma_2 = \|\boldsymbol{\Sigma}_2\| = \|\boldsymbol{M}_1\| &= \max_{\boldsymbol{x}\in\mathbb{C}^{n-1},\, \|\boldsymbol{x}\|=1} \|\boldsymbol{M}_1\boldsymbol{x}\| \\ &\leq \max_{\boldsymbol{x}\in\mathbb{C}^{n},\, \|\boldsymbol{x}\|=1} \left\| \begin{pmatrix} \sigma_1 & \boldsymbol{0}_{1\times(m-1)} \\ \boldsymbol{0}_{(m-1)\times 1} & \boldsymbol{M}_1 \end{pmatrix} \boldsymbol{x} \right\| = \|\boldsymbol{U}_1^*\boldsymbol{M}\boldsymbol{V}_1\| = \|\boldsymbol{M}\| = \sigma_1. \end{aligned}$$

Furthermore, with

$$\boldsymbol{U} = \boldsymbol{U}_1 \begin{pmatrix} 1 & \boldsymbol{0}_{1\times(m-1)} \\ \boldsymbol{0}_{(m-1)\times 1} & \boldsymbol{U}_2 \end{pmatrix}, \quad \boldsymbol{V} = \boldsymbol{V}_1 \begin{pmatrix} 1 & \boldsymbol{0}_{1\times(n-1)} \\ \boldsymbol{0}_{(n-1)\times 1} & \boldsymbol{V}_2 \end{pmatrix}$$

$\boldsymbol{U}$, $\boldsymbol{V}$ are unitary and we have

$$\begin{aligned} \boldsymbol{U}^*\boldsymbol{M}\boldsymbol{V} &= \begin{pmatrix} 1 & \boldsymbol{0}_{1\times(m-1)} \\ \boldsymbol{0}_{(m-1)\times 1} & \boldsymbol{U}_2^* \end{pmatrix} \boldsymbol{U}_1^*\boldsymbol{M}\boldsymbol{V}_1 \begin{pmatrix} 1 & \boldsymbol{0}_{1\times(n-1)} \\ \boldsymbol{0}_{(n-1)\times 1} & \boldsymbol{V}_2 \end{pmatrix} \\ &= \begin{pmatrix} \sigma_1 & \boldsymbol{0}_{1\times(n-1)} \\ \boldsymbol{0}_{(m-1)\times 1} & \boldsymbol{\Sigma}_2 \end{pmatrix}. \end{aligned}$$

Thus, the result follows by induction. □

From the SVD, we can derive an economic and a truncated version, which we will use later on. The following definition introduces notation for these.

Definition 2.1.10 (Economic and truncated SVD)**.**
Let $\boldsymbol{M} \in \mathbb{C}^{m\times n}$ be a matrix with $\operatorname{rank}(\boldsymbol{M}) = r$*. Furthermore, let $s = \min\{m, n\}$ and $\boldsymbol{M} = \boldsymbol{U}_0\boldsymbol{\Sigma}_0\boldsymbol{V}_0^*$ be its SVD. We split $\boldsymbol{U_0}$ and $\boldsymbol{V_0}$ according to the positive singular values. Let $\boldsymbol{\Sigma} := \operatorname{diag}(\sigma_1, \ldots, \sigma_r) \in \mathbb{C}^{r\times r}$ and $\boldsymbol{U} \in \mathbb{C}^{m\times r}$, $\boldsymbol{V} \in \mathbb{C}^{n\times r}$ contain the respective first r left and right singular vectors. Let the matrices $\boldsymbol{U}_\perp \in \mathbb{C}^{m\times(s-r)}$, $\boldsymbol{V}_\perp \in \mathbb{C}^{n\times(s-r)}$ contain the remaining singular vectors, respectively. Then we have*

$$\boldsymbol{M} = \boldsymbol{U}_0\boldsymbol{\Sigma}_0\boldsymbol{V}_0^* = \begin{pmatrix} \boldsymbol{U} & \boldsymbol{U}_\perp \end{pmatrix} \begin{pmatrix} \boldsymbol{\Sigma} & \boldsymbol{0}_{r\times(n-r)} \\ \boldsymbol{0}_{(m-r)\times r} & \boldsymbol{0}_{(m-r)\times(n-r)} \end{pmatrix} \begin{pmatrix} \boldsymbol{V}^* \\ \boldsymbol{V}_\perp^* \end{pmatrix} = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^* \tag{2.1.4}$$

and call the most right hand expression the economic SVD *of $\boldsymbol{M}$. For $1 \leq k \leq r$, let $\boldsymbol{\Sigma}_k := \operatorname{diag}(\sigma_1, \ldots, \sigma_k, 0, \ldots, 0) \in \mathbb{C}^{m\times n}$. Then*

$$\boldsymbol{M}_k := \boldsymbol{U}_0\boldsymbol{\Sigma}_k\boldsymbol{V}_0^* \tag{2.1.5}$$

*is the k-*truncated SVD *of $\boldsymbol{M}$ and we call the matrix $\boldsymbol{M}_k$ the k-*truncated SVD matrix *of $\boldsymbol{M}$.*

Remark 2.1.11 (Best rank-k approximation)**.**
Let $\boldsymbol{M} \in \mathbb{C}^{m\times n}$ with $r = \operatorname{rank}(\boldsymbol{M})$. Then for $k \leq r$, the k-truncated SVD matrix of $\boldsymbol{M}$

is a best rank-k approximation to $\boldsymbol{M}$ *with respect to the spectral norm, i.e. if* $\boldsymbol{M}_k$ *is the k-truncated SVD matrix of* $\boldsymbol{M}$*, then*

$$\min_{\boldsymbol{B}\in\mathbb{C}^{m\times n},\,\mathrm{rank}(\boldsymbol{B})=k} \|\boldsymbol{M}-\boldsymbol{B}\| = \|\boldsymbol{M}-\boldsymbol{M}_k\|\,.$$

This result can be found for instance in [49, Thm. 2.5.3].

Some basic properties of singular values are the following.

Lemma 2.1.12 (Properties of singular values)**.**
Let $\boldsymbol{M},\boldsymbol{N}\in\mathbb{C}^{m\times n}$ *and* $s=\max\{m,n\}$*. Then we have*

i) the singular values are invariant under complex conjugation and transposing, i.e. for $j=1,\dots,s$ *it holds*

$$\sigma_j(\boldsymbol{M})=\sigma_j(\boldsymbol{M}^\top),\quad \sigma_j(\boldsymbol{M})=\sigma_j(\overline{\boldsymbol{M}}),$$

and thus,

$$\sigma_j(\boldsymbol{M})=\sigma_j(\boldsymbol{M}^*).$$

ii) Singular values are unitary invariant, which means for any unitary matrices $\boldsymbol{Q}\in\mathbb{C}^{m\times m}$ *and* $\boldsymbol{Q}'\in\mathbb{C}^{n\times n}$ *it holds*

$$\sigma_j(\boldsymbol{Q}\boldsymbol{M}\boldsymbol{Q}')=\sigma_j(\boldsymbol{M}),\quad j=1,\dots,s.$$

Proof. i) and ii) are simple consequences of the SVD and the fact that the singular values are unique, cf. [54, p. 146]. □

Remark 2.1.13 (Orthonormal bases)**.**
The columns of $\boldsymbol{U}$*,* $\boldsymbol{V}$*,* $\boldsymbol{U}_\perp$ *and* $\boldsymbol{V}_\perp$ *form orthonormal bases for* $\mathrm{range}(\boldsymbol{M})$*,* $\mathrm{range}(\boldsymbol{M}^*)$*,* $\ker(\boldsymbol{M}^*)$ *and* $\ker(\boldsymbol{M})$*, respectively, which follows directly through the SVD, Lemma 2.1.12 and especially by* $\boldsymbol{U}_0$ *and* $\boldsymbol{V}_0$ *being unitary.*

The Singular value decomposition allows the analysis of matrices and related quantities, and it is fundamental for matrix perturbation theory. We summarize some advanced properties of singular values in the following lemma. In particular, the connection of singular values to eigenvalues and the spectral norm are very useful.

Lemma 2.1.14 (Connections between singular values and eigenvalues, cf. [55, pp. 450])**.**
Let $\boldsymbol{M}\in\mathbb{C}^{m\times n}$*,* $r=\mathrm{rank}(\boldsymbol{M})$*,* $s=\min\{m,n\}$ *and* $\sigma_1\geq\cdots\geq\sigma_r>0=\sigma_{r+1}=\cdots=\sigma_s$ *be the singular values of* $\boldsymbol{M}$*. Then we have the following.*

i) Define the Hermitian matrix

$$\boldsymbol{B}=\begin{pmatrix}\boldsymbol{0}_m & \boldsymbol{M}\\ \boldsymbol{M}^* & \boldsymbol{0}_n\end{pmatrix}\in\mathbb{C}^{(m+n)\times(m+n)}.$$

Then the ordered eigenvalues of $\boldsymbol{B}$ *are*

$$-\sigma_1\leq\cdots\leq-\sigma_s\leq\underbrace{0=\cdots=0}_{|m-n|\ \text{times}}\leq\sigma_s\leq\cdots\leq\sigma_1.$$

ii) The nonzero eigenvalues of the positive semidefinite matrix $\boldsymbol{M}^*\boldsymbol{M}$ are the squared singular values of $\boldsymbol{M}$, i.e.

$$\{\sigma_j \mid j = 1, \dots, r\} = \left\{ \sqrt{\lambda_j(\boldsymbol{M}^*\boldsymbol{M})} \mid j = 1, \dots, n \right\} \setminus \{0\}.$$

In particular, we have $\sigma_{\max}(\boldsymbol{M}) = \|\boldsymbol{M}\|$.

iii) If $\boldsymbol{M}$ is square and normal, i.e. $\boldsymbol{M}\boldsymbol{M}^* = \boldsymbol{M}^*\boldsymbol{M}$, then for each $\sigma_j, j = 1, \dots, m$ there exists an eigenvalue $\lambda_j(\boldsymbol{M})$ such that

$$\sigma_j = |\lambda_j(\boldsymbol{M})|.$$

Especially, if $\boldsymbol{M}$ is additionally Hermitian and positive semidefinite and the eigenvalues are ordered decreasingly as the singular values are, then

$$\sigma_j = \lambda_j(\boldsymbol{M}),$$

for all $j = 1, \dots, m$.

Proof. i) can be found in [55, Thm. 7.3.3]. A matrix $\boldsymbol{Y} \in \mathbb{C}^{(m+n)\times(m+n)}$ is constructed explicitly such that $\boldsymbol{Y}^*\boldsymbol{B}\boldsymbol{Y}$ is a diagonal matrix with entries $-\sigma_1, \dots, -\sigma_s, \sigma_s, \dots, \sigma_1, 0, \dots, 0$. For the construction the matrices from the SVD of $\boldsymbol{M}$ are used.

ii) is due to the fact that if $\boldsymbol{U}_0\boldsymbol{\Sigma}_0\boldsymbol{V}_0^*$ is the SVD of $\boldsymbol{M}$ then the singular value and eigenvalue decomposition of $\boldsymbol{M}^*\boldsymbol{M}$ coincide and are given by $\boldsymbol{M}^*\boldsymbol{M} = \boldsymbol{V}_0\boldsymbol{\Sigma}_0^*\boldsymbol{\Sigma}_0\boldsymbol{V}_0^*$. The diagonal matrix $\boldsymbol{\Sigma}_0^*\boldsymbol{\Sigma}_0$ has r positive entries $\sigma_1^2, \dots, \sigma_r^2$ which are also the non-zero eigenvalues of $\boldsymbol{M}^*\boldsymbol{M}$ and for these the relation from ii) holds.

In iii) the matrix $\boldsymbol{M}$ is assumed to be square and normal which guarantees that there exists an orthonormal basis of eigenvectors (see [55, Thm. 2.5.3]) which, collected as columns in $\boldsymbol{V}$, yield the eigenvalue decomposition of $\boldsymbol{M} = \boldsymbol{V}^*\boldsymbol{\Lambda}\boldsymbol{V}$. The matrix $\boldsymbol{\Lambda}$ is diagonal with the eigenvalues of $\boldsymbol{M}$ as entries. Using this decomposition, directly yields that $\boldsymbol{M}\boldsymbol{M}^*$ has the same eigenvalues squared. Finally, we apply ii) to obtain the connection between the singular values and the eigenvalues of $\boldsymbol{M}$. Since the eigenvalues of $\boldsymbol{M}$ can have arbitrary signum, equality holds only up to taking the absolute value. This is not the case for $\boldsymbol{M}$ being positive semidefinite. All eigenvalues are non-negative leading to direct correspondence between eigenvalues and singular values. □

Lemma 2.1.14 i) allows to proof further advanced results for singular values by transferring known results about eigenvalues for Hermitian matrices. Similarly to eigenvalues, singular values have a variational characterization.

Theorem 2.1.15 (Variational characterization of singular values, cf. [55, Thm. 7.3.8])**.** *Let $\boldsymbol{M} \in \mathbb{C}^{m\times n}$, $s = \min\{m, n\}$ and $\sigma_1 \geq \cdots \geq \sigma_s$ the singular values of $\boldsymbol{M}$. Let $k \in \{1, \dots, s\}$ and let $\mathcal{S}$ be a subspace of $\mathbb{C}^n$. Then*

$$\sigma_k = \min_{[\mathcal{S}:\dim\mathcal{S}=n-k+1]} \max_{[\boldsymbol{x}\in\mathcal{S},\boldsymbol{x}\neq\boldsymbol{0}]} \frac{\|\boldsymbol{M}\boldsymbol{x}\|}{\|\boldsymbol{x}\|} = \min_{[\mathcal{S}:\dim\mathcal{S}=n-k+1]} \max_{[\boldsymbol{x}\in\mathcal{S},\|\boldsymbol{x}\|=1]} \|\boldsymbol{M}\boldsymbol{x}\| \tag{2.1.6}$$

and

$$\sigma_k = \max_{[\mathcal{S}:\dim\mathcal{S}=k]} \min_{[\boldsymbol{x}\in\mathcal{S},\boldsymbol{x}\neq\boldsymbol{0}]} \frac{\|\boldsymbol{M}\boldsymbol{x}\|}{\|\boldsymbol{x}\|} = \max_{[\mathcal{S}:\dim\mathcal{S}=k]} \min_{[\boldsymbol{x}\in\mathcal{S},\|\boldsymbol{x}\|=1]} \|\boldsymbol{M}\boldsymbol{x}\|. \tag{2.1.7}$$

Particularly, this confirms

$$\sigma_{\max}(\boldsymbol{M}) = \max_{\boldsymbol{x}\in\mathbb{C}^n,\|\boldsymbol{x}\|=1} \|\boldsymbol{M}\boldsymbol{x}\| = \|\boldsymbol{M}\|$$

and we additionally have, if $\operatorname{rank}(\boldsymbol{M}) = n$*, then*

$$\sigma_{\min}(\boldsymbol{M}) = \min_{\boldsymbol{x}\in\mathbb{C}^n,\|\boldsymbol{x}\|=1} \|\boldsymbol{M}\boldsymbol{x}\|\,.$$

Proof. The proof can be found in [55, Thm. 7.3.8] and is based on Theorem 2.1.4 and Lemma 2.1.14. The special cases follow from (2.1.6) for $\sigma_{\max}(\boldsymbol{M})$ and from (2.1.7) for $\sigma_{\min}(\boldsymbol{M})$. □

Lemma 2.1.16 (Weyl's inequality & sub-multiplicativity, cf. [54, Thm. 3.3.16])**.**
Let $\boldsymbol{M},\boldsymbol{N}\in\mathbb{C}^{m\times n}$ *and* $s=\min\{m,n\}$*. The Weyl inequality for singular values*

$$\sigma_{j+k-1}(\boldsymbol{M}+\boldsymbol{N}) \le \sigma_j(\boldsymbol{M}) + \sigma_k(\boldsymbol{N}), \tag{2.1.8}$$

and the sub-multiplicativity

$$\sigma_{j+k-1}(\boldsymbol{M}\boldsymbol{N}^*) \le \sigma_j(\boldsymbol{M})\sigma_k(\boldsymbol{N}) \tag{2.1.9}$$

hold for all $1\le j,k\le s$ *such that* $1\le j+k-1\le s$*. In particular, we have*

$$|\sigma_j(\boldsymbol{M}+\boldsymbol{N}) - \sigma_j(\boldsymbol{M})| \le \sigma_{\max}(\boldsymbol{N})$$

for $j=1,\dots,s$*.*

Furthermore, if $\boldsymbol{N}\in\mathbb{C}^{n\times\ell}$ *then*

$$\sigma_s(\boldsymbol{M})\sigma_j(\boldsymbol{N}) \le \sigma_j(\boldsymbol{M}\boldsymbol{N}) \le \sigma_{\max}(\boldsymbol{M})\sigma_j(\boldsymbol{N}) \tag{2.1.10}$$

for $j=1,\dots,\min\{n,\ell\}$*.*

Proof. The first two inequalities can be found in [54, Thm. 3.3.16]. The special case follows from choosing $k=1$ in (2.1.8). This yields $\sigma_j(\boldsymbol{M}+\boldsymbol{N})-\sigma_j(\boldsymbol{M})\le\sigma_1(\boldsymbol{N})=\sigma_{\max}(\boldsymbol{N})$. Then we repeat this but now for the replaced matrices $\boldsymbol{M}$ with $\boldsymbol{M}+\boldsymbol{N}$ and $\boldsymbol{N}$ with $-\boldsymbol{N}$. Thus, (2.1.8) gives $\sigma_j(\boldsymbol{M}+\boldsymbol{N}-\boldsymbol{N})=\sigma_j(\boldsymbol{M})\le\sigma_j(\boldsymbol{M}+\boldsymbol{N})+\sigma_{\max}(-\boldsymbol{N})=\sigma_j(\boldsymbol{M}+\boldsymbol{N})+\sigma_{\max}(\boldsymbol{N})$. This is equivalent to $\sigma_j(\boldsymbol{M})-\sigma_j(\boldsymbol{M}+\boldsymbol{N})\le\sigma_{\max}(\boldsymbol{N})$. Both inequalities combined yield the inequality for the absolute value. The inequalities in (2.1.10) can be proven with means of Theorem 2.1.15. First of all we observe for any $\boldsymbol{x}\in\mathbb{C}^\ell$ and given the SVD $\boldsymbol{M}=\boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^*$, where $\boldsymbol{U},\boldsymbol{V}$ are unitary and $\boldsymbol{\Sigma}=\operatorname{diag}\sigma_1(\boldsymbol{M}),\cdots,\sigma_s(\boldsymbol{M})\in\mathbb{C}^{m\times n}$,

$$\|\boldsymbol{M}\boldsymbol{N}\boldsymbol{x}\|^2 = \|\boldsymbol{\Sigma}\boldsymbol{V}^*\boldsymbol{N}\boldsymbol{x}\|^2 = \sum_{j=1}^{m}\left|(\boldsymbol{\Sigma}\boldsymbol{V}^*\boldsymbol{N}\boldsymbol{x})_j\right|^2 \ge \sigma_s(\boldsymbol{M})^2\|\boldsymbol{V}^*\boldsymbol{N}\boldsymbol{x}\|^2 = \sigma_s(\boldsymbol{M})^2\|\boldsymbol{N}\boldsymbol{x}\|^2.$$

Analogously, we obtain $\|\boldsymbol{M}\boldsymbol{N}\boldsymbol{x}\|^2 \le \sigma_{\max}(\boldsymbol{M})^2\|\boldsymbol{N}\boldsymbol{x}\|^2$. Applying both inequalities after taking the square root and denoting by $\mathcal{S}$ a subspace of $\mathbb{C}^\ell$ yields

$$\begin{aligned}\sigma_s(\boldsymbol{M})\max_{[\mathcal{S}:\dim\mathcal{S}=k]}\min_{[\boldsymbol{x}\in\mathcal{S},\|\boldsymbol{x}\|=1]}\|\boldsymbol{N}\boldsymbol{x}\| &\le \max_{[\mathcal{S}:\dim\mathcal{S}=k]}\min_{[\boldsymbol{x}\in\mathcal{S},\|\boldsymbol{x}\|=1]}\|\boldsymbol{M}\boldsymbol{N}\boldsymbol{x}\| \\ &\le \sigma_{\max}(\boldsymbol{M})\max_{[\mathcal{S}:\dim\mathcal{S}=k]}\min_{[\boldsymbol{x}\in\mathcal{S},\|\boldsymbol{x}\|=1]}\|\boldsymbol{N}\boldsymbol{x}\|\end{aligned}$$

and by Theorem 2.1.15 the result. □

Lemma 2.1.17 (Hermitian contraction)**.**
Let $\boldsymbol{M} \in \mathbb{C}^{m\times n}$ and $\boldsymbol{H} \in \mathbb{C}^{n\times n}$ be a Hermitian matrix. If $\|\boldsymbol{H}\| \leq 1$, it is called Hermitian contraction *and for the eigenvalues of the Hermitian matrices $\boldsymbol{M}\boldsymbol{M}^*$ and $\boldsymbol{M}\boldsymbol{H}\boldsymbol{M}^*$ we have*

$$\lambda_j(\boldsymbol{M}\boldsymbol{M}^*) \geq \lambda_j(\boldsymbol{M}\boldsymbol{H}\boldsymbol{M}^*)$$

for all $j = 1, \ldots, m$. If additionally $\boldsymbol{H}$ is positive semidefinite, then we also have

$$\sigma_j(\boldsymbol{M}\boldsymbol{M}^*) \geq \sigma_j(\boldsymbol{M}\boldsymbol{H}\boldsymbol{M}^*)$$

for all $j = 1, \ldots, m$.

Proof. [55, Thm. 7.7.2 a)] tells us, if $\boldsymbol{G} \in \mathbb{C}^{n\times n}$ is a second Hermitian matrix and $\boldsymbol{G} - \boldsymbol{H}$ is positive semidefinite, then also $\boldsymbol{M}\boldsymbol{G}\boldsymbol{M}^* - \boldsymbol{M}\boldsymbol{H}\boldsymbol{M}^*$ is positive semidefinite. Now, in our lemma $\boldsymbol{G}$ is replaced by the identity matrix $\boldsymbol{I}_n$ which is clearly Hermitian. If $\boldsymbol{V}\boldsymbol{\Lambda}\boldsymbol{V}$ is a unitary eigenvalue decomposition for $\boldsymbol{H}$, then $\boldsymbol{V}^*(\boldsymbol{I}_n - \boldsymbol{\Lambda})\boldsymbol{V}$ is an eigenvalue decomposition for $\boldsymbol{I}_n - \boldsymbol{H}$. Furthermore, the eigenvalues of $\boldsymbol{I}_n - \boldsymbol{H}$ are then explicitly given by $1 - \lambda_j(\boldsymbol{H}), j = 1, \ldots, n$ and since $\|\boldsymbol{H}\| \leq 1$ implies $\lambda_{\max}(\boldsymbol{H}) \leq 1$, we have $\boldsymbol{I}_n - \boldsymbol{H}$ is positive semidefinite. Hence, $\boldsymbol{M}\boldsymbol{M}^* - \boldsymbol{M}\boldsymbol{H}\boldsymbol{M}^*$ is positive semidefinite. By [55, Cor. 7.7.4 c)] we learn that the eigenvalues of two Hermitian matrices $\boldsymbol{H}_1, \boldsymbol{H}_2 \in \mathbb{C}^{m\times m}$ satisfy the inequalities $\lambda_j(\boldsymbol{H}_1) \geq \lambda_j(\boldsymbol{H}_2)$ for all j if their difference $\boldsymbol{H}_1 - \boldsymbol{H}_2$ is positive semidefinite. Applied to the above yields the inequalities for the eigenvalues.

The inequalities for the singular values follow, because Lemma 2.1.14 tells that they coincide with the eigenvalues in the positive semidefinite case. The matrix $\boldsymbol{M}\boldsymbol{M}^*$ is positive semidefinite by construction and $\boldsymbol{M}\boldsymbol{H}\boldsymbol{M}^*$ is positive semidefinite if the middle matrix $\boldsymbol{H}$ is positive semidefinite. This follows by using Theorem 2.1.4 for the smallest eigenvalue and interpreting $\boldsymbol{M}^*\boldsymbol{x}$ as test vectors, or can alternatively be found in [55, Observation 7.1.8 a)]. □

Lemma 2.1.18 (Interlacing property of singular values, cf. [55, Cor. 7.3.6])**.**
Let $\boldsymbol{M} \in \mathbb{C}^{m\times n}$, $s = \min\{m, n\}$ and $\widetilde{\boldsymbol{M}}$ be the matrix obtained from $\boldsymbol{M}$ by deleting one row or column. Let $\sigma_1 \geq \cdots \geq \sigma_s$ be the singular values of $\boldsymbol{M}$ and $\widetilde{\sigma}_1 \geq \cdots \geq \widetilde{\sigma}_s$ the ones of $\widetilde{\boldsymbol{M}}$, where $\widetilde{\sigma}_s = 0$ if $m \geq n$ and a column is deleted, or if $n \geq m$ and a row is deleted. Then we have the interlacing property of the singular values

$$\sigma_1 \geq \widetilde{\sigma}_1 \geq \sigma_2 \geq \widetilde{\sigma}_2 \geq \cdots \geq \widetilde{\sigma}_{s-1} \geq \sigma_s \geq \widetilde{\sigma}_s.$$

In particular, we have the monotony

$$\|\boldsymbol{M}\| = \sigma_{\max}(\boldsymbol{M}) \geq \sigma_{\max}(\widetilde{\boldsymbol{M}}) = \left\|\widetilde{\boldsymbol{M}}\right\|$$

and if $\operatorname{rank}(\boldsymbol{M}) = \operatorname{rank}(\widetilde{\boldsymbol{M}})$*, then also the monotony*

$$\sigma_{\min}(\boldsymbol{M}) \geq \sigma_{\min}(\widetilde{\boldsymbol{M}}).$$

Proof. This is a consequence of Lemma 2.1.14 i) combined with Lemma 2.1.6. The inequalities for the $\sigma_{\max}$ are already included, and since $\sigma_{\min}$ is the smallest non-zero singular value, the equality of the ranks ensures that the corresponding indices of $\sigma_{\min}(\boldsymbol{M})$ and $\sigma_{\min}(\widetilde{\boldsymbol{M}})$ are the same in the above chain of inequalities. □

Special norm bounds and norm equalities

Lemma 2.1.19 (Matrices with orthonormal columns)**.**
Let $\boldsymbol{U} = [\boldsymbol{u}_1, \dots, \boldsymbol{u}_n]$ *and* $\boldsymbol{V} = [\boldsymbol{v}_1, \dots, \boldsymbol{v}_n] \in \mathbb{C}^{m\times n}$ *with* $\operatorname{range}(\boldsymbol{U}) = \operatorname{range}(\boldsymbol{V})$ *and where the columns* $\boldsymbol{u}_j$ *and* $\boldsymbol{v}_j$ *are orthonormal bases for* $\operatorname{range}(\boldsymbol{U})$*, respectively. Then we have*

i)

$$\|\boldsymbol{U}\| = 1,$$

ii) for any $\boldsymbol{y} \in \mathbb{C}^n$ *and* $\boldsymbol{Y} \in \mathbb{C}^{n\times \ell}$

$$\|\boldsymbol{U}\boldsymbol{y}\| = \|\boldsymbol{y}\|, \quad \|\boldsymbol{U}\boldsymbol{Y}\| = \|\boldsymbol{Y}\|,$$

iii) and there exists a unique, unitary matrix $\boldsymbol{R} \in \mathbb{C}^{n\times n}$*, such that*

$$\boldsymbol{V} = \boldsymbol{U}\boldsymbol{R}.$$

Proof. Let $\boldsymbol{y} = (y_1, \dots, y_n)^\top \in \mathbb{C}^n$. Then we have

$$\|\boldsymbol{U}\boldsymbol{y}\|^2 = \left\langle \sum_{k=1}^n y_k \boldsymbol{u}_k, \sum_{\ell=1}^n y_\ell \boldsymbol{u}_\ell \right\rangle = \sum_{k=1}^n \sum_{\ell=1}^n y_k \overline{y}_\ell \left\langle \boldsymbol{u}_k, \boldsymbol{u}_\ell \right\rangle = \sum_{k=1}^n y_k \overline{y}_k = \|\boldsymbol{y}\|^2.$$

This directly shows the first part in ii) after taking the square root. Dividing by $\|\boldsymbol{y}\|^2$ and taking the maximum over all such vectors $\boldsymbol{y}$ we have $\|\boldsymbol{U}\|^2 = 1$ by Definition 2.1.7 i). The second part of ii) follows by applying the first part of ii) in

$$\|\boldsymbol{U}\boldsymbol{Y}\| = \max_{\boldsymbol{y}\in\mathbb{C}^n,\, \|\boldsymbol{y}\|=1} \|\boldsymbol{U}(\boldsymbol{Y}\boldsymbol{y})\| = \max_{\boldsymbol{y}\in\mathbb{C}^n,\, \|\boldsymbol{y}\|=1} \|\boldsymbol{Y}\boldsymbol{y}\| = \|\boldsymbol{Y}\|.$$

For iii) we do the following. Since $\operatorname{range}(\boldsymbol{U}) = \operatorname{range}(\boldsymbol{V})$ and the columns build bases, respectively, we have for each $j = 1, \dots, n$ that there exists a unique coefficient vector $\boldsymbol{r}_j \in \mathbb{C}^d$, such that $\boldsymbol{v}_j = \boldsymbol{U}\boldsymbol{r}_j$. Therefore, with $\boldsymbol{R} = [\boldsymbol{r}_1, \dots, \boldsymbol{r}_n]$, we obtain

$$\boldsymbol{V} = \boldsymbol{U}\boldsymbol{R}.$$

Finally, the orthonormality of the columns of $\boldsymbol{U}$ and $\boldsymbol{V}$ yields

$$\boldsymbol{I}_n = \boldsymbol{V}^*\boldsymbol{V} = \boldsymbol{R}^*\boldsymbol{U}^*\boldsymbol{U}\boldsymbol{R} = \boldsymbol{R}^*\boldsymbol{R},$$

which shows $\boldsymbol{R}$ is unitary. □

The following lemma exploits the properties of Hermitian positive definite matrices to bound the norm of the inverse in terms of the matrix itself.

Lemma 2.1.20 (Norm of matrix inverse)**.**
Let $\boldsymbol{M} \in \mathbb{C}^{m\times m}$ *Hermitian and positive definite and* $\boldsymbol{I}_m \in \mathbb{C}^{m\times m}$ *the identity matrix. Let* $\eta \in \mathbb{R}$ *be a parameter satisfying* $\eta > \|\boldsymbol{M}\|$*, then*

$$\left\|\boldsymbol{M}^{-1}\right\| = \frac{1}{\eta - \|\eta \boldsymbol{I}_m - \boldsymbol{M}\|}.$$

Proof. We use Lemma 2.1.14. Firstly, observe $\boldsymbol{M}$ is positive definite and therefore, there exists an unitary eigenvector matrix $\boldsymbol{V}$, such that $\boldsymbol{M} = \boldsymbol{V}\operatorname{diag}(\lambda_1(\boldsymbol{M}), \dots, \lambda_m(\boldsymbol{M}))\boldsymbol{V}^*$ with real positive eigenvalues $\lambda_1(\boldsymbol{M}) \geq \cdots \geq \lambda_m(\boldsymbol{M}) > 0$. The computation

$$\begin{aligned}\eta\boldsymbol{I}_m - \boldsymbol{M} &= \boldsymbol{V}\left(\eta\boldsymbol{V}^*\boldsymbol{V} - \operatorname{diag}(\lambda_1(\boldsymbol{M}), \dots, \lambda_m(\boldsymbol{M}))\right)\boldsymbol{V}^* \\ &= \boldsymbol{V}\operatorname{diag}(\eta - \lambda_1(\boldsymbol{M}), \dots, \eta - \lambda_m(\boldsymbol{M}))\boldsymbol{V}^*\end{aligned}$$

shows that the eigenvalues of $\eta\boldsymbol{I}_m - \boldsymbol{M}$ are given by $\eta - \lambda_j(\boldsymbol{M})$, for $j = 1, \dots, m$. Furthermore, since by assumption $\eta > \|\boldsymbol{M}\| = \lambda_{\max}(\boldsymbol{M})$, all such eigenvalues are positive and hence, can be ordered by $\eta - \lambda_{\min}(\boldsymbol{M}) = \eta - \lambda_m(\boldsymbol{M}) \geq \cdots \geq \eta - \lambda_1(\boldsymbol{M}) = \eta - \lambda_{\max}(\boldsymbol{M})$. This finally leads to

$$\left\|\boldsymbol{M}^{-1}\right\| = \frac{1}{\lambda_{\min}(\boldsymbol{M})} = \frac{1}{\eta - (\eta - \lambda_{\min}(\boldsymbol{M}))} = \frac{1}{\eta - \lambda_{\max}(\eta\boldsymbol{I}_m - \boldsymbol{M})} = \frac{1}{\eta - \|\eta\boldsymbol{I}_m - \boldsymbol{M}\|}.$$

□

Lemma 2.1.21 (Schur-complement decomposition, cf. [55, eq. (0.8.5.3)])**.**
Let $\ell, m \in \mathbb{N}$ and $\boldsymbol{M} \in \mathbb{C}^{(\ell+m)\times(\ell+m)}$ be a 2×2 block matrix of the form

$$\boldsymbol{M} = \begin{pmatrix}\boldsymbol{M}_1 & \boldsymbol{M}_2 \\ \boldsymbol{M}_3 & \boldsymbol{M}_4\end{pmatrix}, \quad \boldsymbol{M}_1 \in \mathbb{C}^{\ell\times\ell},\, \boldsymbol{M}_4 \in \mathbb{C}^{m\times m},$$

with $\boldsymbol{M}_1$ being invertible. Then the Schur-complement decomposition *is given by*

$$\boldsymbol{M} = \begin{pmatrix}\boldsymbol{I}_\ell & \boldsymbol{0} \\ -\boldsymbol{M}_3\boldsymbol{M}_1^{-1} & \boldsymbol{I}_m\end{pmatrix}^{-1}\begin{pmatrix}\boldsymbol{M}_1 & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{M}_4 - \boldsymbol{M}_3\boldsymbol{M}_1^{-1}\boldsymbol{M}_2\end{pmatrix}\begin{pmatrix}\boldsymbol{I}_\ell & -\boldsymbol{M}_1^{-1}\boldsymbol{M}_2 \\ \boldsymbol{0} & \boldsymbol{I}_m\end{pmatrix}^{-1}.$$

The block $[\boldsymbol{M}/\boldsymbol{M}_1] := \boldsymbol{M}_4 - \boldsymbol{M}_3\boldsymbol{M}_1^{-1}\boldsymbol{M}_2 \in \mathbb{C}^{m\times m}$ is called Schur-complement *of $\boldsymbol{M}_1$ in $\boldsymbol{M}$.*

Proof. We simply revert the multiplication by inverted matrices and perform block matrix multiplication on the left hand side to obtain the block diagonal matrix. □

Lemma 2.1.22 (Block Gerschgorin theorem, cf. [55, 6.1.P17] or [40, Thm. 5])**.**
Let $\boldsymbol{M} \in \mathbb{C}^{nm\times nm}$ be an $m \times m$ block matrix with blocks $\boldsymbol{M}_{ik} \in \mathbb{C}^{n\times n}$. Let the diagonal blocks $\boldsymbol{M}_{ii}$ be normal ($\boldsymbol{M}^*\boldsymbol{M} = \boldsymbol{M}\boldsymbol{M}^*$) and denote $\lambda_1^{(i)}, \dots, \lambda_n^{(i)}$ their respective eigenvalues. Then the eigenvalues of $\boldsymbol{M}$ are included in the set

$$\bigcup_{i=1}^{n}\bigcup_{j=1}^{m}\left\{z \in \mathbb{C}\colon \left|z - \lambda_j^{(i)}\right| \leq \sum_{k=1,\,k\neq i}^{m}\|\boldsymbol{M}_{ik}\|\right\}.$$

In particular, if $n = 1$, we have the statement from the classical Gerschgorin theorem, cf. [55, Thm. 6.1.1]. Furthermore, we have for $\boldsymbol{M} \in \mathbb{C}^{m\times n}$ the inequalities

$$\left\|\begin{pmatrix}\boldsymbol{0}_{n\times n} & \boldsymbol{M}^* \\ \boldsymbol{M} & \boldsymbol{0}_{m\times m}\end{pmatrix}\right\| \leq \|\boldsymbol{M}\| \quad \textit{and} \quad \left\|\begin{pmatrix}\boldsymbol{I}_n & \boldsymbol{0}_{n\times m} \\ \boldsymbol{M} & \boldsymbol{I}_m\end{pmatrix}\right\|^2 \leq 1 + \|\boldsymbol{M}\| + \|\boldsymbol{M}\|^2.$$

Proof. The proof can be found in the references, we only explain why the special cases are true. The first matrix is Hermitian and therefore, the absolute values of its eigenvalues match with its singular values. If $m = n$, the Block Gerschgorin theorem directly yields that all eigenvalues of the matrix are in a circle with radius $\|\boldsymbol{M}\|$ around zero. Thus, all eigenvalues have absolute value smaller than $\|\boldsymbol{M}\|$. If $m > n$ or $n > m$, then adding zero rows and columns ensures that we have square blocks and can apply the Block Gerschgorin theorem.

For the second inequality, we first need the observation that $\|\boldsymbol{N}\|^2 = \|\boldsymbol{N}^*\boldsymbol{N}\|$ for any matrix, which follows simply by Lemma 2.1.14. Then we have

$$\left\| \begin{pmatrix} \boldsymbol{I}_n & \boldsymbol{0}_{n\times m} \\ \boldsymbol{M} & \boldsymbol{I}_m \end{pmatrix} \right\|^2 = \left\| \begin{pmatrix} \boldsymbol{I}_n + \boldsymbol{M}^*\boldsymbol{M} & \boldsymbol{M}^* \\ \boldsymbol{M} & \boldsymbol{I}_m \end{pmatrix} \right\|$$

and proceed analogously to the first inequality. □

The following result can be found in [103, p. 53] without proof and similar estimates for the Frobenius norm are given in [55, p. 520].

Lemma 2.1.23 (Matrix norm monotonicity with respect to entries, cf. [103, 55])**.**
Let $\boldsymbol{M}, \boldsymbol{N} \in \mathbb{C}^{m\times n}$ with entries satisfying $\left|(\boldsymbol{M})_{k,\ell}\right| \leq (\boldsymbol{N})_{k,\ell}$ for all $k = 1, \dots, m$, $\ell = 1, \dots, n$, then

$$\|\boldsymbol{M}\| \leq \|\boldsymbol{N}\|.$$

Proof. Let $\boldsymbol{x} = (x_1, \dots, x_n)^\top \in \mathbb{C}^n$. We directly show the result by

$$\begin{aligned} \|\boldsymbol{M}\|^2 &= \max_{\|\boldsymbol{x}\|=1} \|\boldsymbol{M}\boldsymbol{x}\|^2 = \max_{\|\boldsymbol{x}\|=1} \sum_{k=1}^{m} \left| \sum_{\ell=1}^{n} (\boldsymbol{M})_{k,\ell}\, x_\ell \right|^2 \leq \max_{\|\boldsymbol{x}\|=1} \sum_{k=1}^{m} \left(\sum_{\ell=1}^{n} \left|(\boldsymbol{M})_{k,\ell}\right| |x_\ell| \right)^2 \\ &\leq \max_{\|\boldsymbol{x}\|=1} \sum_{k=1}^{m} \left(\sum_{\ell=1}^{n} (\boldsymbol{N})_{k,\ell}\, |x_\ell| \right)^2 = \max_{\|\boldsymbol{x}\|=1} \sum_{k=1}^{m} \left(\sum_{\ell=1}^{n} (\boldsymbol{N})_{k,\ell}\, x_\ell \right)^2 = \|\boldsymbol{N}\|^2. \end{aligned}$$

□

The condition number

Before we come to the definition of the condition number, we start with the well-known Moore-Penrose pseudo inverse.

Definition 2.1.24 (Moore-Penrose pseudo inverse)**.**
Let $\boldsymbol{M} \in \mathbb{C}^{m\times n}$ with $\operatorname{rank}(\boldsymbol{M}) = r$ *and SVD $\boldsymbol{M} = \boldsymbol{U}_0\boldsymbol{\Sigma}_0\boldsymbol{V}_0^*$ with diagonal matrix $\boldsymbol{\Sigma}_0 = \operatorname{diag}(\sigma_1, \dots, \sigma_r, 0, \dots, 0) \in \mathbb{C}^{m\times n}$. Then the* Moore-Penrose pseudo inverse *is given by*

$$\boldsymbol{M}^\dagger := \boldsymbol{V}_0\boldsymbol{\Sigma}_0^\dagger\boldsymbol{U}_0^* \in \mathbb{C}^{n\times m} \quad \text{with} \quad \boldsymbol{\Sigma}_0^\dagger := \operatorname{diag}(\sigma_1^{-1}, \dots, \sigma_r^{-1}, 0, \dots, 0) \in \mathbb{C}^{n\times m}.$$

Lemma 2.1.25 (Properties of the Moore-Penrose pseudo inverse, cf. [104, Ch. III, Thm. 1.2])**.**
Let $\boldsymbol{M} \in \mathbb{C}^{m\times n}$. Then we have

i) $(\boldsymbol{M}^\dagger)^\dagger = \boldsymbol{M}, \quad (\overline{\boldsymbol{M}})^\dagger = \overline{(\boldsymbol{M}^\dagger)}, \quad (\boldsymbol{M}^\top)^\dagger = (\boldsymbol{M}^\dagger)^\top, \quad (\boldsymbol{M}^*)^\dagger = (\boldsymbol{M}^\dagger)^*.$

ii) Let $\boldsymbol{M}$ have full rank.
If $n = m$, then $\boldsymbol{M}^\dagger = \boldsymbol{M}^{-1}$,
If $n > m$, then $\boldsymbol{M}^\dagger = \boldsymbol{M}^*(\boldsymbol{M}\boldsymbol{M}^*)^{-1}$ is a right inverse, i.e. $\boldsymbol{M}\boldsymbol{M}^\dagger = \boldsymbol{I}_m$
If $n < m$, then $\boldsymbol{M}^\dagger = (\boldsymbol{M}^*\boldsymbol{M})^{-1}\boldsymbol{M}^*$ is a left inverse, i.e. $\boldsymbol{M}^\dagger\boldsymbol{M} = \boldsymbol{I}_n$.

iii) Let $\boldsymbol{N} \in \mathbb{C}^{n\times\ell}$. *Then* $(\boldsymbol{MN})^\dagger = \boldsymbol{N}^\dagger\boldsymbol{M}^\dagger$ *only if*
$\boldsymbol{M}$ *has full column rank or*
$\boldsymbol{N}$ *has full row rank or*
$\boldsymbol{N} = \boldsymbol{M}^*$.

With the pseudo inverse at hand, we can motivate and define the condition number. Let $\boldsymbol{Mx} = \boldsymbol{b}$, with $\boldsymbol{M} \in \mathbb{C}^{m\times n}$, $m \geq n$ and $\operatorname{rank}(\boldsymbol{M}) = n$. The solution is $\boldsymbol{x} = \boldsymbol{M}^\dagger\boldsymbol{b}$ and we consider the perturbed system $\boldsymbol{M}\tilde{\boldsymbol{x}} = \tilde{\boldsymbol{b}}$, with solution $\tilde{\boldsymbol{x}} = \boldsymbol{M}^\dagger\tilde{\boldsymbol{b}}$ in the least squares sense. Then the relative error between the solutions are bounded by

$$\frac{\|\tilde{\boldsymbol{x}}\|}{\|\boldsymbol{x}\|} = \frac{\left\|\boldsymbol{M}^\dagger\tilde{\boldsymbol{b}}\right\|}{\|\boldsymbol{x}\|}\cdot\frac{\|\boldsymbol{b}\|}{\|\boldsymbol{b}\|} \leq \frac{\left\|\boldsymbol{M}^\dagger\right\|\left\|\tilde{\boldsymbol{b}}\right\|\|\boldsymbol{M}\|\,\|\boldsymbol{x}\|}{\|\boldsymbol{x}\|\,\|\boldsymbol{b}\|} = \|\boldsymbol{M}\|\left\|\boldsymbol{M}^\dagger\right\|\frac{\left\|\tilde{\boldsymbol{b}}\right\|}{\|\boldsymbol{b}\|}$$

This motivates the definition of the condition number of the system matrix $\boldsymbol{M}$. In our example it is the amplification factor for the relative error which is independent of the chosen solution algorithm.

Definition 2.1.26 (Condition number)**.**
The spectral condition number *of a matrix* $\boldsymbol{M} \in \mathbb{C}^{m\times n}$ *with full rank is defined by*

$$\operatorname{cond}(\boldsymbol{M}) := \|\boldsymbol{M}\|\left\|\boldsymbol{M}^\dagger\right\|,$$

which is due to Lemma 2.1.14 ii) and Definition 2.1.24 equal to

$$\operatorname{cond}(\boldsymbol{M}) = \frac{\sigma_{\max}(\boldsymbol{M})}{\sigma_{\min}(\boldsymbol{M})}.$$

Sometimes different norms than the spectral norm are used to define the condition number. In this case we give the condition number the respective index, for instance $\operatorname{cond}_\infty := \|\boldsymbol{M}\|_\infty\left\|\boldsymbol{M}^\dagger\right\|_\infty$ *and we note that these condition numbers are mutually equivalent as the norms are by Lemma 2.1.8.*

The following lemma provides two properties of the spectral condition number

Lemma 2.1.27.
Let $\boldsymbol{M} \in \mathbb{C}^{m\times n}$ *have full rank, then we always have*

$$\operatorname{cond}(\boldsymbol{M}) \geq 1.$$

Furthermore,

$$\operatorname{cond}(\boldsymbol{M}) = 1$$

if and only if $\boldsymbol{MM}^* = a\boldsymbol{I}_m$ *and* $m \leq n$ *or* $\boldsymbol{M}^*\boldsymbol{M} = a\boldsymbol{I}_n$ *and* $m \geq n$ *for some* $a \in \mathbb{R}_{>0}$.

Proof. $\operatorname{cond}(\boldsymbol{M}) \geq 1$ is obvious since $\sigma_{\min}(\boldsymbol{M}) \leq \sigma_{\max}(\boldsymbol{M})$. Without loss of generality we assume that $m \leq n$, otherwise we use that $\operatorname{cond}(\boldsymbol{M}) = \operatorname{cond}(\boldsymbol{M}^*)$ as the spectral norm is unitary invariant. Let $\boldsymbol{U\Sigma V}^*$ be the SVD of $\boldsymbol{M}$. If $\boldsymbol{MM}^* = a\boldsymbol{I}$ for some $a > 0$, then we have, using Lemma 2.1.14, $\sigma_{\min}(\boldsymbol{M}) = \sqrt{\lambda_{\min}(\boldsymbol{MM^*})} = \sqrt{\lambda_{\max}(\boldsymbol{MM}^*)} = \sigma_{\max}(\boldsymbol{M})$ and thus, $\operatorname{cond}(\boldsymbol{M}) = 1$. Now we assume $\boldsymbol{M}$ has full rank m and $\operatorname{cond}(\boldsymbol{M}) = 1$. Hence, we necessarily have $\sigma_{\max}(\boldsymbol{M}) = \sigma_{\min}(\boldsymbol{M})$ and $\sigma_1(\boldsymbol{M}) = \cdots = \sigma_m(\boldsymbol{M})$. Using the SVD of $\boldsymbol{M}$ we obtain

$$\boldsymbol{MM}^* = \boldsymbol{U\Sigma V}^*\boldsymbol{V\Sigma}^*\boldsymbol{U}^* = \boldsymbol{U\Sigma\Sigma}^*\boldsymbol{U}^*,$$

where $\boldsymbol{\Sigma}\boldsymbol{\Sigma}^* = \operatorname{diag}(\sigma_1(\boldsymbol{M})^2, \ldots, \sigma_1(\boldsymbol{M})^2) \in \mathbb{C}^{m\times m}$. Let $\boldsymbol{u}_1, \ldots, \boldsymbol{u}_m \in \mathbb{C}^m$ denote the row vectors of $\boldsymbol{U}$. Then applying that $\boldsymbol{U}$ is unitary yields for each entry $1 \leq j, k \leq m$

$$(\boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{\Sigma}^*\boldsymbol{U}^*)_{j,k} = \sum_{\ell=1}^{m} (\boldsymbol{u}_j)_\ell \, \sigma_1(\boldsymbol{M})^2 \overline{(\boldsymbol{u}_k)_\ell} = \sigma_1(\boldsymbol{M})^2 \langle \boldsymbol{u}_j, \boldsymbol{u}_k \rangle = \sigma_1(\boldsymbol{M})^2 \delta_{j,k},$$

where $\delta_{j,k}$ denotes the Kronecker delta. This completes the proof. □

The condition number and extremal singular values appear in many matrix perturbation results that are given in the next section.

2.2 Matrix perturbation theory

In this section, we collect matrix perturbation results which are useful tools for analyzing the stability of subspace methods, see Chapter 4. A corollary of the Bauer-Fike Theorem allows to control the deviation of eigenvalues when a diagonalizable matrix is perturbed. In connection with the singular value decomposition, a theorem by Wedin shows how sensitive subspaces spanned by singular vectors behave when matrix entries are perturbed. For that, principal angles and vectors allow to measure the distance between two subspaces. Additionally, a theorem also by Wedin enables us to measure the sensitivity of the pseudo inverse on perturbations of the inverted matrix.

We start with defining the matching distance between two discrete sets of the same cardinality.

Definition 2.2.1 (Matching distance, cf. [104, Ch. 4 Def.1.2])**.**
Let $\Lambda = \{\lambda_1, \ldots, \lambda_m\}, \widetilde{\Lambda} = \{\widetilde{\lambda}_1, \ldots, \widetilde{\lambda}_m\} \subset \mathbb{C}$ *and* $\pi\colon \{1, \ldots, m\} \to \{1, \ldots, m\}$ *denote a permutation. Then the* matching distance *between* Λ *and* $\widetilde{\Lambda}$ *is defined by*

$$\operatorname{md}(\Lambda, \widetilde{\Lambda}) := \min_{\pi} \max_{j\in\{1,\ldots,m\}} \left|\lambda_j - \widetilde{\lambda}_{\pi(j)}\right|.$$

Remark 2.2.2 (Hausdorff distance, cf. [104, pp. 167])**.**
Let Λ *and* $\widetilde{\Lambda}$ *from Definition 2.2.1. Then the* Hausdorff distance *is defined by*

$$\operatorname{hd}(\Lambda, \widetilde{\Lambda}) := \max\left\{ \max_{[1\leq j\leq m]} \min_{[1\leq k\leq m]} \left|\widetilde{\lambda}_j - \lambda_k\right|, \max_{[1\leq k\leq m]} \min_{[1\leq j\leq m]} \left|\widetilde{\lambda}_j - \lambda_k\right| \right\}.$$

It is the well-known Hausdorff distance applied to discrete sets. Furthermore, the matching distance is at least as large as the Hausdorff distance. For instance the sets $\{1, 3, 11\}$ *and* $\{2, 10, 13\}$ *have matching distance* 7 *and Hausdorff distance* 2*. If the matching distance is small, it is ensured, that each element of the first set has exactly one element nearby from the second set. Thus, the matching distance is most appropriate for measuring the displacement of eigenvalues when matrices are perturbed.*

The following theorem is a corollary of the Bauer-Fike Theorem and allows to bound the matching distance between eigenvalues of a diagonalizable matrix and its perturbed version in terms of the perturbation size and the condition number of the diagonalizing matrix.

Theorem 2.2.3 ([104, Ch. 4, Thm. 3.3])**.**
Let $\boldsymbol{M} \in \mathbb{C}^{m\times m}$ be diagonalizable, i.e. there exists a regular matrix $\boldsymbol{P} \in \mathbb{C}^{m\times m}$ such that $\boldsymbol{P}^{-1}\boldsymbol{M}\boldsymbol{P}$ is diagonal. Let $\widetilde{\boldsymbol{M}} = \boldsymbol{M} + \boldsymbol{E}$ and let Λ and $\widetilde{\Lambda}$ denote the set of the respective m eigenvalues of $\boldsymbol{M}$ and $\widetilde{\boldsymbol{M}}$. Then we have

$$\operatorname{md}(\Lambda, \widetilde{\Lambda}) \leq (2m-1)\operatorname{cond}(\boldsymbol{P})\,\|\boldsymbol{E}\|\,.$$

In the following we describe principal angles and principal vectors and their properties. They help to describe the relation between two different subspaces of same dimension in $\mathbb{C}^m$ for some $m \in \mathbb{N}_+$ and how close they are to each other. In particular, if the largest principal angle is zero, we can conclude that both subspaces are the same.

Definition 2.2.4 (Principal angles & principal vectors, cf. [21])**.**
Let $\mathcal{U}, \mathcal{V} \subset \mathbb{C}^m$ be d_1 and d_2 dimensional subspaces and $s = \min\{d_1, d_2\}$. Then the real principal angles *between them, $0 \leq \theta_1 \leq \cdots \leq \theta_s \leq \frac{\pi}{2}$, are iteratively defined by*

$$\cos(\theta_j) := \max_{\left[\substack{\boldsymbol{u}\in\mathcal{U},\,\|\boldsymbol{u}\|=1,\\ \langle\boldsymbol{u},\boldsymbol{u}_k\rangle=0,\,k=1,\ldots,j-1}\right]} \max_{\left[\substack{\boldsymbol{v}\in\mathcal{V},\,\|\boldsymbol{v}\|=1,\\ \langle\boldsymbol{v},\boldsymbol{v}_k\rangle=0,\,k=1,\ldots,j-1}\right]} |\langle\boldsymbol{u},\boldsymbol{v}\rangle| =: |\langle\boldsymbol{u}_j,\boldsymbol{v}_j\rangle|\,,$$

for $j = 1,\ldots,s$. The principal vectors *are given by $\{\boldsymbol{u}_1,\ldots,\boldsymbol{u}_s\} \subset \mathcal{U}$ and $\{\boldsymbol{v}_1,\ldots,\boldsymbol{v}_s\} \subset \mathcal{V}$. For any $\boldsymbol{z} \in \mathbb{C}^d$ with $\|\boldsymbol{z}\| = 1$ and any $\varphi \in \mathbb{R}$, we have that $\boldsymbol{z}\mathrm{e}^{i\varphi}$ is in the same subspace as $\boldsymbol{z}$ and $\left\|\boldsymbol{z}\mathrm{e}^{i\varphi}\right\| = 1$. Therefore, even if all principal angels are different, principal vectors are not unique since their sign can change. We choose the principal vectors such that $\operatorname{Re}\langle\boldsymbol{u}_j,\boldsymbol{v}_j\rangle = |\langle\boldsymbol{u}_j,\boldsymbol{v}_j\rangle|\,, j = 1,\ldots,s$.*

The following lemma provides a characterization for singular values and singular vectors, which we use to proof the next lemma for principal angles and principal vectors.

Lemma 2.2.5 (Characterization of singular values and vectors, cf. [21, Thm. 1])**.**
Let $\boldsymbol{M} \in \mathbb{C}^{m\times n}$ and $s = \min\{m,n\}$. Then its singular values and vectors are iteratively given by the following. For $j = 1,\ldots,s$ we set

$$\sigma_j = \max_{\left[\substack{\boldsymbol{u}\in\mathbb{C}^m,\,\|\boldsymbol{u}\|=1\\ \langle\boldsymbol{u},\boldsymbol{u}_k\rangle=0,\,k=1,\ldots,j-1}\right]} \max_{\left[\substack{\boldsymbol{v}\in\mathbb{C}^n,\,\|\boldsymbol{v}\|=1\\ \langle\boldsymbol{v},\boldsymbol{v}_k\rangle=0,\,k=1,\ldots,j-1}\right]} |\boldsymbol{u}^*\boldsymbol{M}\boldsymbol{v}| = \left|\boldsymbol{u}_j^*\boldsymbol{M}\boldsymbol{v}_j\right|.$$

For $j > s$, we set $\sigma_j = 0$ and if $m > s$ or $n > s$, we choose the remaining $\boldsymbol{u}_j$ or $\boldsymbol{v}_j$ orthonormal to the respective previously constructed vectors.

Proof. Let $\boldsymbol{u}_1,\ldots,\boldsymbol{u}_m$ be the left singular vectors and $\boldsymbol{v}_1,\ldots,\boldsymbol{v}_n$ be the right singular vectors of $\boldsymbol{M}$. We start to show that the first singular value σ_1 is given by the above characterization. On the one hand, the SVD provides $\boldsymbol{M}\boldsymbol{v}_1 = \sigma_1\boldsymbol{u}_1$ and by properties of the maximum, we obtain

$$\max_{[\boldsymbol{u}\in\mathbb{C}^m,\,\|\boldsymbol{u}\|=1]} \max_{[\boldsymbol{v}\in\mathbb{C}^n,\,\|\boldsymbol{v}\|=1]} |\boldsymbol{u}^*\boldsymbol{M}\boldsymbol{v}| \geq |\boldsymbol{u}_1^*\boldsymbol{M}\boldsymbol{v}_1| = \sigma_1.$$

On the other hand, the Cauchy-Schwarz inequality and the consistency of the spectral norm and the vector 2-norm yields

$$\max_{[\boldsymbol{u}\in\mathbb{C}^m,\,\|\boldsymbol{u}\|=1]} \max_{[\boldsymbol{v}\in\mathbb{C}^n,\,\|\boldsymbol{v}\|=1]} |\boldsymbol{u}^*\boldsymbol{M}\boldsymbol{v}| \leq \max_{[\boldsymbol{u}\in\mathbb{C}^m,\,\|\boldsymbol{u}\|=1]} \max_{[\boldsymbol{v}\in\mathbb{C}^n,\,\|\boldsymbol{v}\|=1]} \|\boldsymbol{u}\|\,\|\boldsymbol{M}\|\,\|\boldsymbol{v}\| = \|\boldsymbol{M}\| = \sigma_1.$$

Together, we obtain equality. The identity for the next singular value σ_2 can be shown as follows. The lower bound is obtained analogously by using the second singular vectors. For the upper bound, we use that the singular value decomposition provides the rank-1 decomposition

$$\boldsymbol{M} = \sum_{j=1}^{r} \sigma_j \boldsymbol{u}_j \boldsymbol{v}_j^*,$$

where $r = \operatorname{rank}(\boldsymbol{M})$ (the matrices $\boldsymbol{u}_j \boldsymbol{v}_j^*$ have rank one). Using this, again the Cauchy-Schwarz inequality and the consistency of matrix and vector 2-norms, we obtain

$$\begin{aligned}\max_{\substack{\boldsymbol{u}\in\mathbb{C}^m,\,\|\boldsymbol{u}\|=1\\ \langle \boldsymbol{u},\boldsymbol{u}_1\rangle=0}} \max_{\substack{\boldsymbol{v}\in\mathbb{C}^n,\,\|\boldsymbol{v}\|=1\\ \langle \boldsymbol{v},\boldsymbol{v}_1\rangle=0}} |\boldsymbol{u}^*\boldsymbol{M}\boldsymbol{v}| &= \max_{\substack{\boldsymbol{u}\in\mathbb{C}^m,\,\|\boldsymbol{u}\|=1\\ \langle \boldsymbol{u},\boldsymbol{u}_1\rangle=0}} \max_{\substack{\boldsymbol{v}\in\mathbb{C}^n,\,\|\boldsymbol{v}\|=1\\ \langle \boldsymbol{v},\boldsymbol{v}_1\rangle=0}} \left| \sum_{j=1}^{r} \sigma_j \boldsymbol{u}^*\boldsymbol{u}_j\boldsymbol{v}_j^*\boldsymbol{v} \right| \\ = \max_{\substack{\boldsymbol{u}\in\mathbb{C}^m,\,\|\boldsymbol{u}\|=1\\ \langle \boldsymbol{u},\boldsymbol{u}_1\rangle=0}} \max_{\substack{\boldsymbol{v}\in\mathbb{C}^n,\,\|\boldsymbol{v}\|=1\\ \langle \boldsymbol{v},\boldsymbol{v}_1\rangle=0}} \left| \boldsymbol{u} \left(\sum_{j=2}^{r} \sigma_j \boldsymbol{u}_j \boldsymbol{v}_j^* \right) \boldsymbol{v}^* \right| &\leq \left\| \sum_{j=2}^{r} \sigma_j \boldsymbol{u}_j \boldsymbol{v}_j^* \right\|.\end{aligned}$$

Since the last sum can be written as $[\boldsymbol{u}_2, \cdots, \boldsymbol{u}_m] \operatorname{diag}(\sigma_2, \ldots, \sigma_r)[\boldsymbol{v}_2, \cdots, \boldsymbol{v}_n]^*$, we get by sub-multiplicativity and Lemma 2.1.19

$$\max_{\substack{\boldsymbol{u}\in\mathbb{C}^m,\,\|\boldsymbol{u}\|=1\\ \langle \boldsymbol{u},\boldsymbol{u}_1\rangle=0}} \max_{\substack{\boldsymbol{v}\in\mathbb{C}^n,\,\|\boldsymbol{v}\|=1\\ \langle \boldsymbol{v},\boldsymbol{v}_1\rangle=0}} |\boldsymbol{u}^*\boldsymbol{M}\boldsymbol{v}| \leq \sigma_2.$$

The identities for the remaining singular values follow analogously. □

Now we are able to provide a lemma that shows, principal angles and vectors can be computed via SVD.

Theorem 2.2.6 (Computation of principal angles and vectors, cf. [21, Thm. 1])**.**
Let $\boldsymbol{X} \in \mathbb{C}^{m\times n}$ and $\boldsymbol{Y} \in \mathbb{C}^{m\times \ell}$ be arbitrary matrices with columns that form orthonormal bases for the respective ranges $\operatorname{range}(\boldsymbol{X})$ *and* $\operatorname{range}(\boldsymbol{Y})$*. Set $s = \min\{n, \ell\}$,*

$$\boldsymbol{M} = \boldsymbol{X}^*\boldsymbol{Y} \in \mathbb{C}^{n\times \ell}$$

and let the SVD of this matrix be

$$\boldsymbol{M} = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^*.$$

Then the principal angles between $\operatorname{range}(\boldsymbol{X})$ *and* $\operatorname{range}(\boldsymbol{Y})$ *are given by*

$$\theta_j = \arccos(\sigma_j(\boldsymbol{M})), \quad j = 1, \ldots, s.$$

and principal vectors are the columns of the matrices

$$\boldsymbol{X}\boldsymbol{U} \quad \textit{and} \quad \boldsymbol{Y}\boldsymbol{V}. \tag{2.2.1}$$

Proof. Denoting by $\boldsymbol{u}_j$ and $\boldsymbol{v}_j$ the columns of $\boldsymbol{U}$ and $\boldsymbol{V}$, Lemma 2.2.5 yields

$$\sigma_j = \max_{\|\boldsymbol{u}\|=\|\boldsymbol{v}\|=1} |\boldsymbol{u}^*\boldsymbol{M}\boldsymbol{v}| = \left|\boldsymbol{u}_j^*\boldsymbol{M}\boldsymbol{v}_j\right| \tag{2.2.2}$$

subject to $\boldsymbol{u}^*\boldsymbol{u}_k = \boldsymbol{v}^*\boldsymbol{v}_k = 0$ for $k = 1, \dots, j-1$. For $\boldsymbol{x} \in \operatorname{range}(\boldsymbol{X})$ and $\boldsymbol{y} \in \operatorname{range}(\boldsymbol{Y})$ we write

$$\boldsymbol{x} = \boldsymbol{X}\boldsymbol{u} \in \operatorname{range}(\boldsymbol{X}), \quad \boldsymbol{y} = \boldsymbol{Y}\boldsymbol{v} \in \operatorname{range}(\boldsymbol{Y})$$

with suitable vectors $\boldsymbol{u} \in \mathbb{C}^n$ and $\boldsymbol{v} \in \mathbb{C}^\ell$ and analogously the indexed vectors. Then we have due to Lemma 2.1.19 $\|\boldsymbol{x}\| = \|\boldsymbol{u}\|$, $\|\boldsymbol{y}\| = \|\boldsymbol{v}\|$,

$$\boldsymbol{x}^*\boldsymbol{x}_k = \boldsymbol{u}^*\boldsymbol{X}^*\boldsymbol{X}\boldsymbol{u}_k = \boldsymbol{u}^*\boldsymbol{u}_k, \quad \text{and} \quad \boldsymbol{y}^*\boldsymbol{y}_k = \boldsymbol{v}^*\boldsymbol{Y}^*\boldsymbol{Y}\boldsymbol{v}_k = \boldsymbol{v}^*\boldsymbol{v}_k.$$

Furthermore, $\boldsymbol{u}^*\boldsymbol{M}\boldsymbol{v} = \boldsymbol{u}^*\boldsymbol{X}^*\boldsymbol{Y}\boldsymbol{v} = \boldsymbol{x}^*\boldsymbol{y}$ and thus, (2.2.2) is equivalent to

$$\sigma_j = \max_{\|\boldsymbol{x}\|=\|\boldsymbol{y}\|=1} |\boldsymbol{x}^*\boldsymbol{y}| = \left|\boldsymbol{x}_j^*\boldsymbol{y}_j\right|$$

subject to $\boldsymbol{x}^*\boldsymbol{x}_k = \boldsymbol{y}^*\boldsymbol{y}_k = 0$ for $k = 1, \dots, j-1$, which coincides with Definition 2.2.4 and therefore, completes the proof. □

Corollary 2.2.7 (Orthogonality of principal vectors)**.**
Let $\boldsymbol{u}_j, \boldsymbol{v}_j$, for $j = 1, \dots, d$, be the principal vectors between two subspaces constructed by Theorem 2.2.6 and collected as columns in the matrices $\boldsymbol{W} = [\boldsymbol{u}_1, \dots, \boldsymbol{u}_d]$ and $\widetilde{\boldsymbol{W}} = [\boldsymbol{v}_1, \dots, \boldsymbol{v}_d]$. Then it holds

$$\boldsymbol{W}^*\widetilde{\boldsymbol{W}} = \operatorname{diag}(\cos(\theta_1), \dots, \cos(\theta_d)).$$

In particular the principal vectors are mutually orthogonal, i.e.

$$\langle \boldsymbol{u}_j, \boldsymbol{v}_k \rangle = 0,$$

for $j \neq k$.

Proof. By definition of the principal angles and principal vectors it is clear that the diagonal entries of $\boldsymbol{W}^*\widetilde{\boldsymbol{W}}$ are the cosines of the principal angles. We use Theorem 2.2.6 to show that the off-diagonal entries are zero and therefore, these principal vectors are mutually orthogonal. We have $\boldsymbol{W} = \boldsymbol{X}\boldsymbol{U}$ and $\widetilde{\boldsymbol{W}} = \boldsymbol{Y}\boldsymbol{V}$ with matrices $\boldsymbol{X}$ and $\boldsymbol{Y}$ being orthonormal bases for the respective subspaces. The unitary matrices $\boldsymbol{U}$ and $\boldsymbol{V}$ are from the SVD $\boldsymbol{X}^*\boldsymbol{Y} = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^*$. The diagonal matrix $\boldsymbol{\Sigma}$ has the cosines of the principal angles on its diagonal. Therefore,

$$\boldsymbol{W}^*\widetilde{\boldsymbol{W}} = \boldsymbol{U}^*\boldsymbol{X}^*\boldsymbol{Y}\boldsymbol{V} = \boldsymbol{U}^*\boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^*\boldsymbol{V} = \boldsymbol{\Sigma}$$

concludes the proof. □

Now we state a lemma which relates the largest principal angle between two subspaces to the norm of the difference of principal vector matrices. In particular this bound is independent of the dimensions and improves over the estimate used in [74, p. 13].

Lemma 2.2.8 (Subspace difference)**.**
Let $n \leq m$ and $\boldsymbol{W}, \widetilde{\boldsymbol{W}} \in \mathbb{C}^{m \times n}$ be matrices with the respective principal vectors of the subspaces $\operatorname{range}(\boldsymbol{M}), \operatorname{range}(\boldsymbol{N}) \subset \mathbb{C}^m$ *as columns. Let $\theta \in [0, \pi/2]$ be the largest principal angle between* $\operatorname{range}(\boldsymbol{M})$ *and* $\operatorname{range}(\boldsymbol{N})$*. Then we have*

$$\left\|\boldsymbol{W} - \widetilde{\boldsymbol{W}}\right\| = 2\sin\left(\frac{\theta}{2}\right) \leq \sqrt{2}\sin(\theta),$$

where the constant $\sqrt{2}$ is best possible for inequality leading to a constant multiple of $\sin(\theta)$*.*

Proof. Let $\theta_1 \leq \cdots \leq \theta_n$ be the principal angles. We use Corollary 2.2.7 to obtain

$$\left\|\boldsymbol{W} - \widetilde{\boldsymbol{W}}\right\|^2 = \left\|(\boldsymbol{W} - \widetilde{\boldsymbol{W}})^*(\boldsymbol{W} - \widetilde{\boldsymbol{W}})\right\| = \left\|\boldsymbol{W}^*\boldsymbol{W} - \widetilde{\boldsymbol{W}}^*\boldsymbol{W} - \boldsymbol{W}^*\widetilde{\boldsymbol{W}} + \widetilde{\boldsymbol{W}}^*\widetilde{\boldsymbol{W}}\right\|$$
$$= 2\left\|\boldsymbol{I}_n - \operatorname{diag}(\cos(\theta_1), \ldots, \cos(\theta_n)\right\| = 2(1 - \cos(\theta_n)).$$

Using the double angle formula $1 - \cos(\theta) = 2\sin(\theta/2)^2$ and taking the square root yields the identity from the lemma. Alternatively, since $0 \leq \cos(\theta) \leq 1$ for $\theta \in \left[0, \frac{\pi}{2}\right]$, we have

$$2(1 - \cos(\theta_n)) \leq 2(1 - \cos(\theta_n)^2) = 2\sin(\theta)^2.$$

Taking the square root yields the inequality. If $\theta = \pi/2$, then

$$2\sin\left(\frac{\theta}{2}\right) = \sqrt{2} = \sqrt{2}\sin(\theta),$$

showing that $\sqrt{2}$ is best possible in the inequality. □

The following theorem by Wedin [112] provides a bound on the perturbation of a subspace spanned by some leading singular vectors of a matrix. The perturbation is measured in terms of the largest principal angle between this subspace and the one spanned by the corresponding singular vectors of the perturbed matrix. This theorem is also known as $\sin(\boldsymbol{\theta})$-theorem in the literature.

Theorem 2.2.9 (Wedin, cf. [112, p. 102])**.**
Let $m, n, k \in \mathbb{N}$ and $\boldsymbol{M}$, $\widetilde{\boldsymbol{M}} = \boldsymbol{M} + \boldsymbol{E} \in \mathbb{C}^{m\times n}$ with $k \leq \operatorname{rank}(\widetilde{\boldsymbol{M}})$. Consider the economic versions of the k-truncated SVD matrices $\boldsymbol{M}_k = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^*$ and $\widetilde{\boldsymbol{M}}_k = \widetilde{\boldsymbol{U}}\widetilde{\boldsymbol{\Sigma}}\widetilde{\boldsymbol{V}}^*$, i.e. $\boldsymbol{U}, \widetilde{\boldsymbol{U}}, \boldsymbol{V}$ and $\widetilde{\boldsymbol{V}}$ consist of the first k singular vectors, respectively. We denote by $\theta_j, \phi_j, j = 1, \ldots k$ the principal angles for the subspace pairs $\operatorname{range}(\boldsymbol{U})$, $\operatorname{range}(\widetilde{\boldsymbol{U}})$ *and* $\operatorname{range}(\boldsymbol{V})$, $\operatorname{range}(\widetilde{\boldsymbol{V}})$*. Let*

$$\sin(\boldsymbol{\theta}) := \operatorname{diag}(\sin(\theta_1), \ldots, \sin(\theta_k)) \quad \textit{and} \quad \sin(\boldsymbol{\phi}) := \operatorname{diag}(\sin(\phi_1), \ldots, \sin(\phi_k)).$$

If there exist $\beta, \gamma > 0$ such that

$$\sigma_k(\boldsymbol{M}) \geq \beta + \gamma \quad \textit{and} \quad \sigma_{k+1}(\widetilde{\boldsymbol{M}}) \leq \beta,$$

then

$$\max\left\{\|\sin(\boldsymbol{\theta})\|, \|\sin(\boldsymbol{\phi})\|\right\} \leq \frac{\max\left\{\|\boldsymbol{E}\boldsymbol{V}\|, \|\boldsymbol{U}^*\boldsymbol{E}\|\right\}}{\gamma}.$$

The spectral norm can be replaced by any other unitary invariant norm.

Remark 2.2.10.
In the above theorem the roles of perturbed and unperturbed matrices is exchanged compared to the [112, p. 102]. This way it is easier to apply in Chapter 4.

Finally, the next theorem, also by Wedin, bounds the perturbation of the Moore-Penrose pseudo inverse of a matrix $\boldsymbol{M}$ in terms of the perturbation of the matrix itself.

Theorem 2.2.11 (Pert. bound for pseudo inverses, [104, Ch. 3 Thm. 3.9], [113, Thm. 4.1])**.**
Let $\boldsymbol{M}, \widetilde{\boldsymbol{M}} = \boldsymbol{M} + \boldsymbol{E} \in \mathbb{C}^{m\times n}$ with $m \geq n$ and $\operatorname{rank}(\boldsymbol{M}) = \operatorname{rank}(\widetilde{\boldsymbol{M}})$, then

$$\left\|\boldsymbol{M}^\dagger - \widetilde{\boldsymbol{M}}^\dagger\right\| \leq \frac{\eta\,\|\boldsymbol{E}\|}{\sigma_{\min}(\boldsymbol{M})\,\sigma_{\min}(\widetilde{\boldsymbol{M}})},$$

where η takes values from the table

$\operatorname{rank}(\boldsymbol{M}) < n$	$\operatorname{rank}(\boldsymbol{M}) = n < m$	$\operatorname{rank}(\boldsymbol{M}) = n = m$
$\frac{1+\sqrt{5}}{2}$	$\sqrt{2}$	1

2.3 Fourier analytic tools

This section serves for introducing notation around function spaces and Fourier analytic tools that are used throughout the thesis. The results and definitions are standard in the literature and only a small part of the Lebesgue integration and Fourier theory. The books [97, 51, 60, 89] are excellent references for detailed presentations of these fields.

We denote by $\mathbb{T} = \mathbb{R}/\mathbb{Z} = [0,1)$ the periodic interval, which parametrizes the complex unit circle by

$$\{z \in \mathbb{C} \mid |z| = 1\} = \left\{\mathrm{e}^{2\pi i t} \mid t \in \mathbb{T}\right\}.$$

For a dimension $d \in \mathbb{N}_+$, the d-fold Cartesian product of $\mathbb{T}$ is given by $\mathbb{T}^d$. Given a multi-index $\boldsymbol{\nu} = (\nu_1, \ldots, \nu_d)^\top \in \mathbb{Z}^d$ and a vector $\boldsymbol{z} = (z_1, \ldots, z_d)^\top \in \mathbb{C}^d$ we define $\boldsymbol{z}^{\boldsymbol{\nu}} := z_1^{\nu_1} \cdot \ldots \cdot z_d^{\nu_d}$ as usual. We use standard notation of well-known related function spaces.

Definition 2.3.1 (Function spaces, cf. [97, Ch. 3], [89, Sec. 1.2])**.**
Let $\mathcal{S}$ be $\mathbb{R}^d$ or $\mathbb{T}^d$. The space of continuous functions *on $\mathcal{S}$ is given by*

$$\mathcal{C}(\mathcal{S}) := \{f \colon \mathcal{S} \to \mathbb{C} \mid f \text{ continuous}\}$$

equipped with the norm

$$\|f\|_{\mathcal{C}(\mathcal{S})} := \sup\{|f(\boldsymbol{t})| \mid \boldsymbol{t} \in \mathcal{S}\}.$$

The Lebesgue spaces of absolute integrable functions *and* square integrable functions *on $\mathcal{S}$ are denoted by*

$$L^1(\mathcal{S}) := \left\{f \colon \mathcal{S} \to \mathbb{C} \mid \|f\|_{L^1(\mathcal{S})} < \infty\right\}$$

with norm $\|f\|_{L^1(\mathcal{S})} := \int_{\mathcal{S}} |f(\boldsymbol{t})| \, d\boldsymbol{t}$ and

$$L^2(\mathcal{S}) := \left\{f \colon \mathcal{S} \to \mathbb{C} \mid \|f\|_{L^2(\mathcal{S})} < \infty\right\}$$

with norm $\|f\|_{L^2(\mathcal{S})} := \left(\int_{\mathcal{S}} |f(\boldsymbol{t})|^2 \, d\boldsymbol{t}\right)^{\frac{1}{2}}$. Furthermore, $L^2(\mathcal{S})$ is a Hilbert space with scalar product

$$\langle \cdot, \cdot \rangle_{L^2(\mathcal{S})} : L^2(\mathcal{S}) \times L^2(\mathcal{S}) \to \mathbb{C}, \quad \langle f, g \rangle_{L^2(\mathcal{S})} := \int_{\mathcal{S}} f(\boldsymbol{t}) \overline{g(\boldsymbol{t})} \, d\boldsymbol{t}.$$

Since $\mathbb{T}^d$ is a compact set, we have the inclusions

$$\mathcal{C}(\mathbb{T}^d) \subset L^2(\mathbb{T}^d) \subset L^1(\mathbb{T}^d), \tag{2.3.1}$$

which follows by the Hölder inequality for general L^p-norms on function spaces, cf. [89, p. 7]. Furthermore, the functions

$$\mathbb{T}^d \to \mathbb{C}, \quad \boldsymbol{t} \mapsto \mathrm{e}^{2\pi i \boldsymbol{\nu}^* \boldsymbol{t}}, \quad \boldsymbol{\nu} \in \mathbb{Z}^d, \tag{2.3.2}$$

build an orthonormal basis for $L^2(\mathbb{T}^d)$ with respect to the L^2-scalar product. A special class of functions in $\mathcal{C}(\mathbb{T}^d)$ are trigonometric polynomials.

Definition 2.3.2 (Trigonometric polynomial, cf. [89, (1.11)]).
A function $f \in \mathcal{C}(\mathbb{T}^d)$ of the form

$$f\colon \mathbb{T}^d \to \mathbb{C}, \quad \boldsymbol{t} \mapsto \sum_{\boldsymbol{\nu}\in\mathbb{Z}^d} c_{\boldsymbol{\nu}} \mathrm{e}^{2\pi i \boldsymbol{\nu}^* \boldsymbol{t}},$$

with finitely many non-zero summands is called trigonometric polynomial. *The set*

$$\mathcal{P}(n) := \left\{ f \in \mathcal{C}(\mathbb{T}^d) \mid f = \sum_{\boldsymbol{\nu}\in\mathbb{N}^d,\, \|\boldsymbol{\nu}\|_\infty \le n} c_{\boldsymbol{\nu}} \mathrm{e}^{2\pi i \boldsymbol{\nu}^* \boldsymbol{t}} \right\}$$

contains all trigonometric polynomials of (max-)degree at most n.

At some points we make use of the Cauchy–Schwarz inequality for $L^2(\mathbb{T}^d)$.

Lemma 2.3.3 (Cauchy–Schwarz inequality, cf. [97, Thm. 3.5]).
Let $f, g \in L^2(\mathbb{T}^d)$, then the Cauchy–Schwarz inequality

$$\left| \langle f, g \rangle_{L^2(\mathbb{T}^d)} \right| \le \|f\|_{L^2(\mathbb{T}^d)} \, \|g\|_{L^2(\mathbb{T}^d)}$$

holds.

For integrable functions we have the Fourier transform at hand.

Definition 2.3.4 (Fourier transform, cf. [89, Sec. 2.1]).
For a function $f \in L^1(\mathbb{R}^d)$ the Fourier transform *is given by*

$$\mathbb{R}^d \to \mathbb{C}, \quad \widehat{f}(\boldsymbol{\omega}) := \left\langle f, \mathrm{e}^{2\pi i \cdot} \right\rangle_{L^2(\mathbb{R}^d)} = \int_{\mathbb{R}^d} f(\boldsymbol{t}) \mathrm{e}^{-2\pi i \boldsymbol{\omega}^* \boldsymbol{t}} \, d\boldsymbol{t}.$$

At some points in the thesis complex measures appear. They are defined as follows.

Definition 2.3.5 (Discrete measures on $\mathbb{T}^d$, cf. [97, 6.1 and p. 188], [60, pp. 380]).
We denote the complex Borel measures on the Borel sets $\mathcal{B}(\mathbb{T}^d)$ by

$$\mathcal{M}(\mathbb{T}^d) := \left\{ \mu\colon \mathcal{B}(\mathbb{T}^d) \to \mathbb{C} \mid \mu \text{ is a measure} \right\}.$$

The Dirac measure $\delta \in \mathcal{M}(\mathbb{T}^d)$ *is given by*

$$\delta\colon \mathcal{B}(\mathbb{T}^d) \to \mathbb{R}_{\ge 0}, \quad \delta(A) := \begin{cases} 1, & 0 \in A, \\ 0, & 0 \notin A. \end{cases}$$

Therefore, the integral of a continuous function $f \in \mathcal{C}(\mathbb{T}^d)$ with respect to δ is given by

$$\int_{\mathbb{T}^d} f(\boldsymbol{t}) \, d\delta(\boldsymbol{t}) = f(\boldsymbol{0}).$$

Let $\boldsymbol{t} \in \mathbb{T}^d$. Then the shifted Dirac measure is defined by $\delta_{\boldsymbol{t}} := \delta(\cdot - \boldsymbol{t})$. Hence, given a support set $\{\boldsymbol{t}_1, \ldots, \boldsymbol{t}_m\} \subset \mathbb{T}^d$ and coefficients $c_1, \ldots, c_m \in \mathbb{C}$, for some $m \in \mathbb{N}$, the complex discrete measure $\mu \in \mathcal{M}(\mathbb{T}^d)$ is defined by

$$\mu\colon \mathcal{B}(\mathbb{T}^d) \to \mathbb{C}, \quad \mu(A) := \sum_{j=1}^{m} c_j \delta_{\boldsymbol{t}_j}(A),$$

and the integral of a function $f \in \mathcal{C}(\mathbb{T}^d)$ with respect to μ is given by

$$\int_{\mathbb{T}^d} f(\boldsymbol{t})\, d\mu(\boldsymbol{t}) = \sum_{j=1}^{m} c_j f(\boldsymbol{t}_j).$$

Definition 2.3.6 (Fourier coefficients, cf. [60, Sec. 4.1 and p. 414])**.**
Let $f \in L^1(\mathbb{T}^d)$, $\mu \in \mathcal{M}(\mathbb{T}^d)$ and $\boldsymbol{z} = \mathrm{e}^{2\pi i \boldsymbol{t}}, \boldsymbol{t} \in \mathbb{T}^d$, then for an integer vector $\boldsymbol{\nu} \in \mathbb{Z}^d$ the respective Fourier coefficients are defined by

$$\widehat{f}\colon \mathbb{Z}^d \to \mathbb{C}, \quad \widehat{f}(\boldsymbol{\nu}) := \int_{\mathbb{T}^d} f(\boldsymbol{t}) \mathrm{e}^{-2\pi i \boldsymbol{\nu}^* \boldsymbol{t}}\, d\boldsymbol{t}$$

and

$$\widehat{\mu}\colon \mathbb{Z}^d \to \mathbb{C}, \quad \widehat{\mu}(\boldsymbol{\nu}) := \int_{\mathbb{T}^d} \mathrm{e}^{-2\pi i \boldsymbol{\nu}^* \boldsymbol{t}}\, d\mu(\boldsymbol{t}).$$

We directly see that the Fourier coefficients and the coefficients of a trigonometric polynomial coincide since the functions $\mathrm{e}^{2\pi i \boldsymbol{\nu}^* \cdot}, \boldsymbol{\nu} \in \mathbb{Z}^d$, form an orthonormal basis in $L^2(\mathbb{T}^d)$. Two multiplied functions also behave.

Lemma 2.3.7 (Convolution, cf. [60, Sec. 2.1])**.**
Let $f, g \in C(\mathbb{T}^d)$ be two trigonometric polynomials. Then their multiplication translates to a discrete convolution of their Fourier coefficients, i.e., for $\boldsymbol{\nu} \in \mathbb{Z}^d$, we have

$$\widehat{(fg)}(\boldsymbol{\nu}) = \sum_{\boldsymbol{\alpha} \in \mathbb{Z}^d} \widehat{f}(\boldsymbol{\alpha}) \widehat{g}(\boldsymbol{\nu} - \boldsymbol{\alpha}).$$

Theorem 2.3.8 (Parseval's identity, cf. [89, Thm. 4.5])**.**
Let $f \in L^2(\mathbb{T}^d)$. We denote the (infinite) vector of its Fourier coefficients by $\widehat{\boldsymbol{f}} := \left(\widehat{f}\right)_{\boldsymbol{\nu} \in \mathbb{Z}^d}$. Then Parseval's identity is

$$\|f\|_{L^2(\mathbb{T}^d)} = \left\|\widehat{\boldsymbol{f}}\right\|,$$

where with slight abuse of notation $\left\|\widehat{\boldsymbol{f}}\right\| := \left(\sum_{\boldsymbol{\nu} \in \mathbb{Z}^d} \left|\widehat{f}(\boldsymbol{\nu})\right|^2\right)^{\frac{1}{2}}$, similar to the Euclidean norm of finite vectors.

An additional useful tool is the Poisson summation formula. It relates the periodization of an integrable function to the Fourier series that has the samples of the Fourier transform on the integer grid as coefficients.

Lemma 2.3.9 (Poisson summation formula, [51, Prop. 1.4.2])**.**
Let $f \in L^1(\mathbb{R}^d)$ and suppose that for some $\epsilon > 0$ and $C > 0$ the function f and its Fourier transform $\widehat{f}$ satisfy the decay conditions $|f(\boldsymbol{t})| \leq C(1+\|\boldsymbol{t}\|)^{-d-\epsilon}$ and $|\widehat{f}(\boldsymbol{\omega})| \leq C(1+\|\boldsymbol{t}\|)^{-d-\epsilon}$. Then the Poisson summation formula

$$\sum_{\boldsymbol{k} \in \mathbb{Z}^d} f(\boldsymbol{t} + \boldsymbol{k}) = \sum_{\boldsymbol{\nu} \in \mathbb{Z}^d} \widehat{f}(\boldsymbol{\nu}) \mathrm{e}^{2\pi i \boldsymbol{\nu}^* \boldsymbol{t}}$$

holds. The identity holds point-wise for all $\boldsymbol{t} \in \mathbb{R}^d$ and both sums converge absolutely for all $\boldsymbol{t} \in \mathbb{R}^d$. In particular, if $\boldsymbol{t} = 0$, we have

$$\sum_{\boldsymbol{k} \in \mathbb{Z}^d} f(\boldsymbol{k}) = \sum_{\boldsymbol{\nu} \in \mathbb{Z}^d} \widehat{f}(\boldsymbol{\nu}).$$

2.4 Auxiliary functions

In the following we collect some special functions and their analytic properties, we need later on. We start with the well-known Gamma function which is useful to analyze and bound factorial terms. Afterwards, we go on with trigonometric kernel functions.

The Gamma function

The *Gamma function* is given by

$$\Gamma\colon \mathbb{R}_{>0} \to \mathbb{R}, \quad \Gamma(t) := \int_0^\infty \mathrm{e}^x x^{t-1}\, dx. \tag{2.4.1}$$

For more background, we refer to [4, Ch. 1] and [115, Ch. 12]. We collect properties relevant for our purposes in the following lemma. The corresponding references are given in the proof.

Lemma 2.4.1 (Properties of the Gamma function)**.**
Let $n \in \mathbb{N}_+$ and $t \in \mathbb{R}_{>0}$. The Gamma function has the explicit values

$$\Gamma(1) = 1, \quad \Gamma\left(\frac{1}{2}\right) = \sqrt{\pi}. \tag{2.4.2}$$

It fulfills the functional equation

$$\Gamma(t+1) = t \cdot \Gamma(t) \tag{2.4.3}$$

and especially interpolates the factorial, i.e.

$$\Gamma(n+1) = n!. \tag{2.4.4}$$

It is logarithmically convex, i.e. for $x, y \in \mathbb{R}_{>0}$ and $0 < a < 1$ we have

$$\Gamma(ax + (1-a)y) \leq \Gamma(x)^a \Gamma(y)^{1-a}. \tag{2.4.5}$$

The Stirling formula yields

$$\sqrt{2\pi} n^{n+\frac{1}{2}} \mathrm{e}^{-n} < \Gamma(n+1) \leq \mathrm{e} n^{n+\frac{1}{2}} \mathrm{e}^{-n} \tag{2.4.6}$$

and furthermore, we have the Gautschi/Wendel inequalities for a parameter $s \in (0,1)$

$$t^{1-s} < \frac{\Gamma(t+1)}{\Gamma(t+s)} < (t+1)^{1-s}. \tag{2.4.7}$$

Proof. The equations (2.4.2) to (2.4.5) are standard results in the literature and can be found for example in [4, p. 3,7,13] and [115, Ch. 12].

Stirling's formula, which can be found for example in [95], yields

$$\sqrt{2\pi} n^{n+\frac{1}{2}} \mathrm{e}^{-n} \mathrm{e}^{\frac{1}{12n+1}} < \Gamma(n+1) < \sqrt{2\pi} n^{n+\frac{1}{2}} \mathrm{e}^{-n} \mathrm{e}^{\frac{1}{12n}}.$$

Since $\mathrm{e}^{\frac{1}{12n+1}} \geq 1$, we directly get the lower bound in (2.4.6). For $n = 1$, equality holds in the upper bound of (2.4.6), and for $n > 1$, we use $\sqrt{2\pi}\mathrm{e}^{\frac{1}{12n}} \leq \mathrm{e}$ in the upper bound from the latter chain of inequalities. However, $\sqrt{2\pi}\mathrm{e}^{\frac{1}{12n}} \leq \mathrm{e}$ is true if and only if $2\pi \leq \mathrm{e}^{2-\frac{1}{6n}}$, which holds since for $n > 2$, we have $2\pi \leq \mathrm{e}^{\frac{23}{12}}$. This can easily be checked.

The Gautschi/Wendel inequalities can be found in [42] and [114]. □

Extremal functions

Beurling reported in the manuscript "On functions with a spectral gap" which can be found in [19] an extremal function, that was later on referred to as the Beurling function by Selberg in [100, p. 226]. It is given in the following lemma together with its properties and we refer to [100, Ch. 45, Sec. 20] and [110].

Lemma 2.4.2 (Beurling function, [110, 100, 19])**.**
The Beurling function*, see Figure 2.4.1 left, is given by*

$$B\colon \mathbb{R}\to\mathbb{R},\quad B(x) := \begin{cases} 1, & x\in\mathbb{N},\\ -1, & -x\in\mathbb{N}_+,\\ \left(\frac{\sin(\pi x)}{2}\right)^2\left[\frac{2}{x}+\sum_{k=1}^{\infty}\frac{1}{(x-k)^2}-\sum_{k=0}^{\infty}\frac{1}{(x+k)^2}\right], & \text{else,}\end{cases}$$

and has the properties:

i) it is a majorant for the sign function, i.e.

$$B(x)\geq \operatorname{sgn}(x),\quad \text{for all } x\in\mathbb{R},$$

ii) it is extended to the complex plane an entire function of exponential type 2π*,*

iii) its Fourier transform $\widehat{B}$ *is supported in* $[-1,1]$*,*

iv) the integral of its difference to the sign function is

$$\int_{\mathbb{R}} B(x)-\operatorname{sgn}(x)\,dx = 1, \tag{2.4.8}$$

and for this it is extremal in the sense that for each function f *satisfying i), ii) and iii) not being* B*, it holds*

$$\int_{\mathbb{R}} f(x)-\operatorname{sgn}(x)\,dx > 1.$$

We remark that in particular the penultimate item can be found in the above reference of Beurling and note that he stated the function $-B(-x)$, a reflected and inverted version of the function given in Lemma 2.4.2, and therefore, a minorant of the sign function. Furthermore, (2.4.8) is easy to see by using

$$B(x)-\operatorname{sgn}(x)+B(-x)-\operatorname{sgn}(-x) = B(x)-B(-x) = 2\frac{\sin(\pi x)^2}{\pi^2x^2}$$

in

$$\begin{aligned} 2\int_{\mathbb{R}} B(x)-\operatorname{sgn}(x)\,dx &= \int_{\mathbb{R}} B(x)-\operatorname{sgn}(x)\,dx + \int_{\mathbb{R}} B(-x)-\operatorname{sgn}(-x)\,dx\\ &= 2\int_{\mathbb{R}}\frac{\sin(\pi x)^2}{\pi^2x^2}\,dx = 2. \end{aligned}$$

Selberg used the Beurling function to construct a minorant and a majorant for the characteristic function over an interval (cf. [110]) and he used the latter for proving a sharp large sieve. The next lemma provides modified versions that turned out to be highly beneficial for proving upper bounds on the condition number of Vandermonde matrices.

Lemma 2.4.3 (Selberg's minorant and majorant, cf. [100, 110])**.**
Let $\chi_{[a,b]}$ be the characteristic function over the interval $[a,b] \subset \mathbb{R}$ and $q > 0$. Then there exists a function $h\colon \mathbb{R} \to \mathbb{R}$ being a majorant for $\chi_{[a,b]}$ that satisfies $\operatorname{supp}(\widehat{h}) \subset [-q,q]$ *and*

$$\widehat{h}(0) = \int_{\mathbb{R}} h(t)\,dt = b - a + \frac{1}{q}.$$

Furthermore, there exists a function $l\colon \mathbb{R} \to \mathbb{R}$ being a minorant for $\chi_{[a,b]}$ that satisfies $\operatorname{supp}(\widehat{l}) \subset [-q,q]$ *and*

$$\widehat{l}(0) = \int_{\mathbb{R}} l(t)\,dt = b - a - \frac{1}{q}.$$

Both functions are entire when extended to the complex plane.

Proof. Let B be the Beurling function from Lemma 2.4.2. Since the characteristic function over an interval $[a_0,b_0] \subset \mathbb{R}$ can be expressed as

$$\chi_{[a_0,b_0]}(x) = \frac{1}{2}\left[\operatorname{sgn}(x-a_0) + \operatorname{sgn}(b_0-x)\right] \tag{2.4.9}$$

and B is a majorant for the sign function, we have

$$h_0\colon \mathbb{R} \to \mathbb{R}, \quad h_0(x) := \frac{1}{2}\left(B(x-a_0) + B(b_0-x)\right), \tag{2.4.10}$$

is a majorant for $\chi_{[a_0,b_0]}$. Furthermore, the Fourier transform is a linear operation and shifting a function leads to modulation of its Fourier transform, hence the Fourier transform of h_0 has the same support as B, i.e. $[-1,1]$. Analogously, since $-B(-x)$ is a minorant for the sign function, we have that

$$l_0\colon \mathbb{R} \to \mathbb{R}, \quad l_0(x) := -\frac{1}{2}\left(B(a_0-x) + B(x-b_0)\right), \tag{2.4.11}$$

is a minorant for $\chi_{[a_0,b_0]}$ again with Fourier transform supported in $[-1,1]$. Now, by using the translation invariance of the integral over the real line, (2.4.9), (2.4.10) and (2.4.8) we obtain for the majorant

$$\begin{aligned}
\widehat{h_0}(0) &= \int_{\mathbb{R}} h_0(x)\,dx = (b_0-a_0) + \int_{\mathbb{R}} h_0(x)\chi_{[a_0,b_0]}(x)\,dx \\
&= (b_0-a_0) + \frac{1}{2}\left[\int_{\mathbb{R}} B(x-a_0) - \operatorname{sgn}(x-a_0)\,dx + \int_{\mathbb{R}} B(b_0-x) - \operatorname{sgn}(b_0-x)\,dx\right] \\
&= (b_0-a_0) + \int_{\mathbb{R}} B(x) - \operatorname{sgn}(x)\,dx = (b_0-a_0) + 1
\end{aligned} \tag{2.4.12}$$

and analogously for the minorant with (2.4.11) and in addition with the rotational symmetry of the sign function around the origin

$$\begin{aligned}
\widehat{l_0}(0) &= \int_{\mathbb{R}} l_0(x)\,dx = (b_0-a_0) + \int_{\mathbb{R}} l_0(x) - \chi_{[a_0,b_0]}(x)\,dx \\
&= (b_0-a_0) - \frac{1}{2}\left[\int_{\mathbb{R}} B(a_0-x) + \operatorname{sgn}(x-a_0)\,dx + \int_{\mathbb{R}} B(x-b_0) + \operatorname{sgn}(b_0-x)\,dx\right]
\end{aligned}$$

$$= (b_0 - a_0) - \frac{1}{2}\left[\int_{\mathbb{R}} B(a_0 - x) - \operatorname{sgn}(a_0 - x)\,dx + \int_{\mathbb{R}} B(x - b_0) + \operatorname{sgn}(x - b_0)\,dx\right]$$
$$= (b_0 - a_0) - \int_{\mathbb{R}} B(x) - \operatorname{sgn}(x)\,dx = (b_0 - a_0) - 1.$$

In order to obtain that both Fourier transforms are supported in $[-q, q]$, we use the dilation property of the Fourier transform. Therefore, choosing $a_0 = qa, b_0 = qb$, then $h\colon \mathbb{R} \to \mathbb{R}, h(x) := h_0(qx)$, has support in $[a_0, b_0]\frac{1}{q} = [a, b]$ and its Fourier transform

$$\widehat{h}(\omega) = \widehat{h_0(q\,\cdot)}(\omega) = \frac{1}{q}\widehat{h_0}(\frac{\omega}{q})$$

has support in $[-q, q]$. Combining the latter equation with (2.4.12), we directly obtain

$$\widehat{h}(0) = \frac{1}{q}\widehat{h_0}(0) = \frac{1}{q}(b_0 - a_0) - \frac{1}{q} = (b - a) - \frac{1}{q}.$$

Analogously, we can proceed with the minorant $l\colon \mathbb{R} \to \mathbb{R}, l(x) := l_0(qx)$, and obtain its desired properties. □

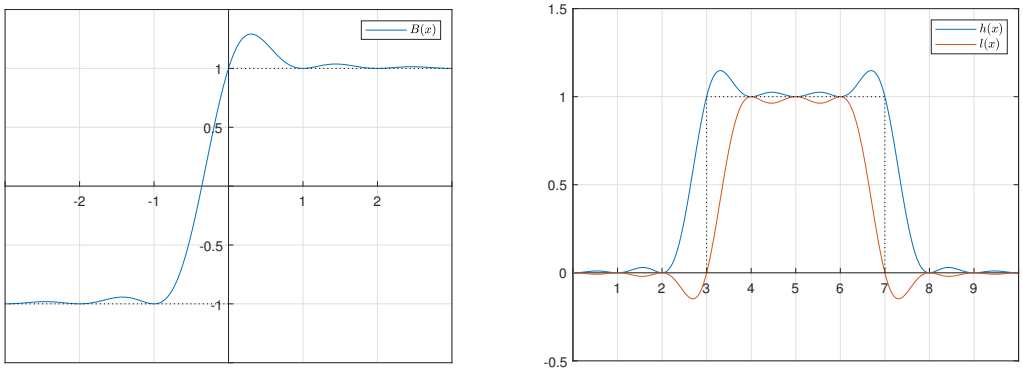


Figure 2.4.1: Left: the Beurling function from Lemma 2.4.2; right: majorant and minorant from Lemma 2.4.3 with $q = 1$ and $[a, b] = [3, 7]$. Involved Beurling functions are processed with the first 1001 summands of each series, respectively.

Remark 2.4.4.
Note, that the minorant from Lemma 2.4.3 is not necessarily zero at the interval endpoints a and b as Figure 2.4.1 suggests. In the figure they are zero since $q = 1$ and $a, b \in \mathbb{N}$ and therefore, in the proof of Lemma 2.4.3 the interval endpoints $a_0, b_0 \in \mathbb{N}$, which, combined with the fact that the Beurling function has values $1, -1$ at integers, leads to this phenomenon. Nevertheless, the minorant from Lemma 2.4.3 is always smaller or equal to zero at the interval endpoints due to being a continuous minorant.

Trigonometric kernel functions

Definition 2.4.5.
Let $n \in \mathbb{N}$ and $N = 2n + 1$. The Dirichlet kernel *of degree n is given by*

$$D_n\colon \mathbb{R} \to \mathbb{R}, \quad D_n(t) := \sum_{k=-n}^{n} \mathrm{e}^{2\pi i k t} = \begin{cases} N, & t \in \mathbb{Z}, \\ \frac{\sin(N\pi t)}{\sin(\pi t)}, & \text{otherwise.} \end{cases} \tag{2.4.13}$$

See Figure 2.4.2 for an example.

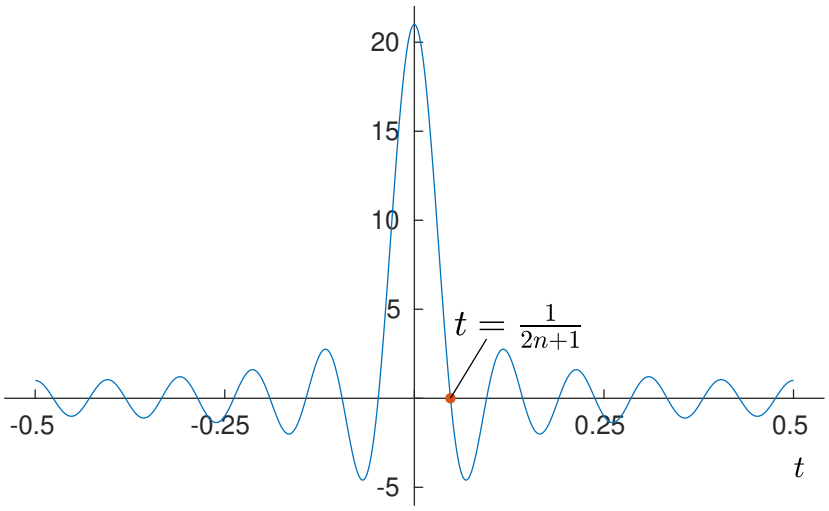


Figure 2.4.2: Dirichlet kernel of degree $n = 10$ over the interval $[-1/2, 1/2]$.

Remark 2.4.6.
The right hand identity for the Dirichlet kernel in (2.4.13) *is obtained by the following. If $t = 0$, we simply plug in the value and resolve the sum. For $t \neq 0$, we use the geometric sum formula and get*

$$\sum_{k=-n}^{n} \mathrm{e}^{2\pi ikt} = \mathrm{e}^{-2\pi int} \sum_{k=0}^{2n} \mathrm{e}^{2\pi ikt} = \mathrm{e}^{-2\pi int} \frac{1-\mathrm{e}^{2\pi i(2n+1)t}}{1-\mathrm{e}^{2\pi it}} = \frac{\mathrm{e}^{-2\pi i(n+\frac{1}{2})t}}{\mathrm{e}^{-2\pi i\frac{t}{2}}} \cdot \frac{1-\mathrm{e}^{2\pi i(2n+1)t}}{1-\mathrm{e}^{2\pi it}}$$
$$= \frac{\mathrm{e}^{-2\pi i(n+\frac{1}{2})t} - \mathrm{e}^{2\pi i(n+\frac{1}{2})t}}{\mathrm{e}^{-2\pi i\frac{t}{2}} - \mathrm{e}^{2\pi i\frac{t}{2}}} = \frac{\sin((2n+1)\pi t)}{\sin(\pi t)}.$$

Lemma 2.4.7.
Let $n \in \mathbb{N}$, $N = 2n+1$. Then the Dirichlet kernel from Definition 2.4.5 is bounded by

$$N - \frac{\pi^2}{6} N^3 t^2 \leq D_n(t) \leq N - N^3 t^2, \qquad 0 \leq |t| \leq \frac{1}{N}.$$

An additional linear lower bound is given by

$$N - N^2 |t| \leq D_n(t), \qquad 0 \leq |t| \leq \frac{1}{N}.$$

Furthermore, the Dirichlet kernel and its first two derivatives are bounded by

$$|D_n(t)| \leq \frac{1}{2|t|},$$
$$|D_n'(t)| \leq N^2 \left(\frac{\pi}{2N|t|} + \frac{1}{2N^2|t|^2} \right),$$
$$|D_n''(t)| \leq N^3 \left(\frac{\pi^2}{2N|t|} + \frac{\pi}{N^2|t|^2} + \frac{1}{N^3|t|^3} \right)$$

for $0 < |t| \leq 1/2$.

Proof. Due to symmetry, it suffices to prove all bounds for $t > 0$ and we use the explicit expression of the Dirichlet kernel in (2.4.13). First, the lower bound on $D_n(t)$ can be derived from the inequalities $x - x^3/6 \leq \sin(x) \leq x$, that hold for all $x \in [0, \pi]$. The left inequality with $x = N\pi t$ and the right inequality with $x = \pi t$ lead to

$$\sin(N\pi t) \geq \left(N - \frac{\pi^2}{6}N^3t^2\right)\pi t \geq \left(N - \frac{\pi^2}{6}N^3t^2\right)\sin(\pi t).$$

The linear lower bound is clearly smaller or equal $N - \frac{\pi^2}{6}N^3t^2$ if $t \leq \frac{1}{2N}$. If $\frac{1}{2N} \leq t \leq \frac{1}{N}$, we show the equivalent inequality

$$\frac{1}{N}D_n\left(\frac{\tilde{t}}{N}\right) = \frac{\sin(\pi\tilde{t})}{N\sin(\pi\tilde{t}/N)} \geq 1 - \tilde{t} \tag{2.4.14}$$

for $\tilde{t} \in [1/2, 1)$. For $\tilde{t} = 1$ obviously equality holds. We use $\sin(x) = \sin(\pi - x)$, for $x \in [\pi/2, \pi]$, and $\pi - x \in [0, \pi/2]$ to obtain $\sin(x) = \sin(\pi - x) \geq \pi - x - (\pi - x)^3/6$. Together with $\sin(x) \leq x$ in (2.4.14) this leads to

$$\begin{aligned}\frac{\sin(\pi\tilde{t})}{N\sin(\pi\tilde{t}/N)} &\geq \frac{1}{\tilde{t}}\left(1 - \tilde{t} - \frac{\pi^2}{6}(1-\tilde{t})^3\right)\\ &\geq (1-\tilde{t})\left(2 - \frac{\pi^2}{3}(1-\tilde{t})^2\right) \geq (1-\tilde{t})\left(2 - \frac{\pi^2}{12}\right) \geq (1-\tilde{t}).\end{aligned}$$

The upper bound on $D_n(t)$ can be derived from the inequality $\cos(\alpha x) \leq \cos(x)$ that holds for all $x \in [0, \pi/2]$ and $\alpha > 1$ such that $\alpha x \in [0, \pi/2]$. Integrating this inequality, choosing $\alpha = N/2$ and $x = \pi t$, and applying the double angle formula yields

$$\frac{\sin(N\pi t)}{2\cos(\frac{N}{2}\pi t)} = \sin(\frac{N}{2}\pi t) \leq \frac{N}{2}\sin(\pi t).$$

Reordering the inequality and applying that $\cos(x) \leq 1 - 4x^2/\pi^2$ for all $x \in [0, \pi/2]$ yields

$$\frac{\sin(N\pi t)}{\sin(\pi t)} \leq N\cos(\frac{N}{2}\pi t) \leq N(1 - N^2t^2).$$

Finally, the remaining bounds on the absolute values can be proven by calculating the first and second derivatives and using $\sin(x) \geq 2x/\pi$ and $\cot x \leq 1/x$ that hold for all $x \in (0, \pi/2]$. □

Lemma 2.4.8 (Trigonometric cancellation, [11, Lem. D.1])**.**
For each $1 \neq z \in \mathbb{C}$ with $|z| = 1$ and for each $m \in \mathbb{N}$ we have

$$\left|\sum_{k=0}^{N} k^m z^k\right| \leq \frac{2}{|1-z|}N^m.$$

Proof. We have

$$(1-z)\sum_{k=0}^{N} k^m z^k = -N^m z^{N+1} + \sum_{k=0}^{N} k^M z^k - \sum_{k=0}^{N-1} k^m z^{k+1} = -N^m z^{N+1} + \sum_{k=1}^{N}(k^m - (k-1)^m)z^k.$$

Thus, by using the triangle inequality we obtain

$$|1-z|\left|\sum_{k=0}^{N} k^m z^k\right| \leq N^m + \sum_{k=1}^{N} k^m - (k-1)^m = 2N^m.$$

Dividing by $|1-z|$ finishes the proof. □

Lemma 2.4.9 (Bounds on Dirichlet kernel derivatives)**.**
Let $0 < t \leq 1/2$, $n \in \mathbb{N}$, $N = 2n+1$ and $\ell \in \mathbb{N}_+$. Then for the derivatives of the Dirichlet kernel we have

$$\left|D_n^{(\ell)}(t)\right| \leq \frac{(2\pi n)^\ell}{t} \leq \frac{(\pi N)^\ell}{t}.$$

Proof. We use the trigonometric cancellation technique used in the proof of [12, Lem. C.1]. First of all we show

$$\sum_{k=1}^{2n} \left|(k-n)^\ell - (k-n-1)^\ell\right| = 2n^\ell. \tag{2.4.15}$$

If ℓ is odd, the differences inside the absolute value are all positive. Hence,

$$\sum_{k=1}^{2n} \left|(k-n)^\ell - (k-n-1)^\ell\right| = \sum_{k=1}^{2n} (k-n)^\ell - (k-n-1)^\ell = -(-n)^\ell + n^\ell = 2n^\ell,$$

where we used that the second term is a telescope sum. If ℓ is even, we split the sum at $k = n, n+1$ and obtain

$$\begin{aligned}
\sum_{k=1}^{2n} \left|(k-n)^\ell - (k-n-1)^\ell\right| &= \sum_{k=1}^{n} \left|(k-n)^\ell - (k-n-1)^\ell\right| + \sum_{k=n+1}^{2n} \left|(k-n)^\ell - (k-n-1)^\ell\right| \\
&= \sum_{k=1}^{n} \left|(n+1-k)^\ell - (n-k)^\ell\right| + \sum_{k=n+1}^{2n} (k-n)^\ell - (k-n-1)^\ell \\
&= \sum_{k=1}^{n} \Big((n+1-k)^\ell - (n-k)^\ell\Big) + n^\ell = 2n^\ell.
\end{aligned}$$

Now let $z = \mathrm{e}^{2\pi i t}$. We can calculate

$$\begin{aligned}
(1-z)\sum_{k=0}^{2n}(k-n)^\ell z^k &= \sum_{k=0}^{2n}(k-n)^\ell z^k - \sum_{k=0}^{2n}(k-n)^\ell z^{k+1} \\
&= \sum_{k=0}^{2n}(k-n)^\ell z^k - \sum_{k=1}^{2n+1}(k-n-1)^\ell z^k \\
&= -n^\ell + \sum_{k=1}^{2n}\Big[(k-n)^\ell - (k-n-1)^\ell\Big] z^k - n^\ell z^{2n+1}.
\end{aligned}$$

Taking the absolute value on both sides, rearranging the equation, applying (2.4.15) and the bound $|1-z| = 2\sin(\pi t) \geq 4t$ yields

$$\left|\sum_{k=0}^{2n}(k-n)^\ell z^k\right| \leq \frac{2n^\ell + \sum_{k=1}^{2n}\left|(k-n)^\ell - (k-n-1)^\ell\right|}{|1-z|} = \frac{4n^\ell}{|1-z|} \leq \frac{n^\ell}{t}. \tag{2.4.16}$$

Finally, we obtain the result by using (2.4.16) in

$$\left|D_n^{(\ell)}(t)\right| = \left|(2\pi i)^\ell \sum_{k=-n}^{n} k^\ell \mathrm{e}^{2\pi i k t}\right| = (2\pi)^\ell \left|\sum_{k=0}^{2n} (k-n)^\ell \mathrm{e}^{2\pi i k t}\right| \leq \frac{(2\pi n)^\ell}{t} \leq \frac{(\pi(2n+1))^\ell}{t}.$$

□

Definition 2.4.10 (Modified Dirichlet kernel)**.**
For $m, \beta \in \mathbb{N}$ the modified Dirichlet kernel *of degree m is defined by $d_m \colon \mathbb{T} \to \mathbb{C}$,*

$$d_m(t) := \frac{1}{m+1} \sum_{k=0}^{m} \mathrm{e}^{2\pi i k t} = \begin{cases} 1, & t = 0, \\ \frac{\mathrm{e}^{\pi i m t}}{m+1} \cdot \frac{\sin(\pi(m+1)t)}{\sin(\pi t)}, & t \neq 0. \end{cases}$$

We define the powers of the multivariate modified Dirichlet kernel by

$$d_m^\beta \colon \mathbb{T}^d \to \mathbb{C}, \quad d_m^\beta(\boldsymbol{t}) := \left(\prod_{\ell=1}^{d} d_m((\boldsymbol{t})_\ell)\right)^\beta \in \mathcal{P}(m\beta).$$

Lemma 2.4.11 (Properties of modified Dirichlet kernels)**.**
Let $m, \beta \in \mathbb{N}_+$ with $m \geq \beta$, $\boldsymbol{\nu} \in \mathbb{Z}^d$ and let d_m^β be a modified Dirichlet kernel of degree m and to the power of β. If $\boldsymbol{t} \in \mathbb{T}^d \setminus \{\mathbf{0}\}$, then

i) $|d_m(\boldsymbol{t})| \leq d_m(\mathbf{0}) = 1$,

ii) $|d_m(\boldsymbol{t})| \leq \frac{1}{2(m+1)|\boldsymbol{t}|_{\mathbb{T}^d}}$,

iii) $\left\|d_m^\beta\right\|_{L^2(\mathbb{T}^d)}^2 \leq \frac{1}{(m+1)^d \beta^{d/2}}$,

iv) $\left|\left\langle d_m^\beta, d_m^\beta(\cdot - \boldsymbol{t})\right\rangle_{L^2(\mathbb{T}^d)}\right| \leq \frac{1}{2(m+1)^d \beta^{(d-1)/2}} \cdot \frac{1}{(m+1)^\beta |\boldsymbol{t}|_{\mathbb{T}^d}^\beta}$,

v) $0 \leq (m+1)^d \widehat{(d_m^\beta)}(\boldsymbol{\nu}) \leq 1$.

Proof. Firstly, note that

$$|d_m(\boldsymbol{t})| \leq \left(\frac{1}{m+1} \sum_{k=0}^{m} \left|\mathrm{e}^{2\pi i k t}\right|\right)^d = 1 = d_m(\mathbf{0})$$

and the point-wise bound follows in the univariate case by

$$|d_m(t)| = \frac{1}{m+1} \left|\frac{\sin(\pi(m+1)t)}{\sin(\pi t)}\right| \leq \frac{1}{(m+1)\,|\sin(\pi t)|} \leq \frac{1}{2(m+1)\,|t|_{\mathbb{T}}}.$$

Secondly, in the multivariate case, setting $t := |\boldsymbol{t}|_{\mathbb{T}^d}$, and using i) and the univariate bound yield

$$|d_m(\boldsymbol{t})| = \prod_{\ell=1}^{d} |d_m((\boldsymbol{t})_\ell)| \leq |d_m(t)| \leq \frac{1}{2(m+1)\,|t|_{\mathbb{T}}} = \frac{1}{2(m+1)\,|\boldsymbol{t}|_{\mathbb{T}^d}}.$$

Note that $\|d_m\|^2_{L^2(\mathbb{T}^d)} = \|d_m\|^{2d}_{L^2(\mathbb{T})}$ and therefore, the third assertion is proven for the univariate case as follows. For $m \geq \beta$, Parseval's identity and direct calculation show

$$\|d_m\|^2_{L^2(\mathbb{T})} = \frac{1}{m+1}, \quad \left\|d_m^2\right\|^2_{L^2(\mathbb{T})} = \frac{1}{m+1}\left[\frac{2}{3} + \frac{1}{3(m+1)^2}\right] \leq \frac{1}{m+1}\cdot\frac{19}{27} \leq \frac{1}{m+1}\cdot\frac{1}{\sqrt{2}},$$
$$\left\|d_m^3\right\|^2_{L^2(\mathbb{T})} = \frac{1}{m+1}\left[\frac{11}{20} + \frac{1}{4(m+1)^2} + \frac{1}{5(m+1)^4}\right] \leq \frac{1}{m+1}\cdot\frac{145}{256} \leq \frac{1}{m+1}\cdot\frac{1}{\sqrt{3}}.$$

For $x \in [0,1]$ and $m \geq 4$, the estimates in [79, Proof of Lemma 2] yield

$$\frac{\sin(\pi x)}{(m+1)\sin\left(\frac{\pi}{m+1}x\right)} \leq \exp\left(-\frac{\pi^2((m+1)^2-1)}{6(m+1)^2}x^2\right) \leq \exp\left(-\frac{4\pi^2x^2}{25}\right)$$

and thus, for $m \geq \beta \geq 4$, the remaining estimate

$$\begin{aligned}
\left\|d_m^\beta\right\|^2_{L^2(\mathbb{T})} &= \frac{2}{(m+1)^{2\beta}}\int_0^{1/2}\left|\frac{\sin(\pi(m+1)t)}{\sin(\pi t)}\right|^{2\beta} dt \\
&= \frac{2}{m+1}\left[\frac{1}{(m+1)^{2\beta}}\left(\int_0^1\left(\frac{\sin(\pi x)}{\sin(\frac{\pi}{m+1}x)}\right)^{2\beta} dx + \int_1^{\frac{m+1}{2}}\left|\frac{\sin(\pi x)}{\sin(\frac{\pi}{m+1}x)}\right|^{2\beta} dx\right)\right] \\
&\leq \frac{2}{m+1}\left[\int_0^\infty \exp\left(-\frac{8\beta\pi^2x^2}{25}\right) dx + \int_1^\infty\left(\frac{1}{2x}\right)^{2\beta} dx\right] \\
&= \frac{1}{m+1}\left[\frac{5}{2\sqrt{2\pi}}\frac{1}{\sqrt{\beta}} + \frac{2^{1-2\beta}}{2\beta-1}\right] \leq \frac{1}{m+1}\cdot\frac{1}{\sqrt{\beta}}.
\end{aligned}$$

In order to prove the fourth assertion, note $|t|_{\mathbb{T}} \leq |t-t'|_{\mathbb{T}} + |t'|_{\mathbb{T}} \leq 2\max\{|t-t'|_{\mathbb{T}}, |t'|_{\mathbb{T}}\}$ and hence, i) and ii) yield

$$\begin{aligned}
\left|d_m(t')\right|\left|d_m(t'-t)\right| &\leq \min\left\{\left|d_m(t'-t)\right|, \left|d_m(t')\right|\right\} \\
&\leq \frac{1}{2(m+1)}\min\left\{\frac{1}{|t'-t|_{\mathbb{T}}}, \frac{1}{|t'|_{\mathbb{T}}}\right\} \leq \frac{1}{(m+1)\,|t|_{\mathbb{T}}}
\end{aligned}$$

and $\|d_m d_m(\cdot - t)\|_{\mathcal{C}(\mathbb{T})} \leq ((m+1)\,|t|_{\mathbb{T}})^{-1}$. Moreover, direct computation gives

$$\begin{aligned}
\left|\langle d_m, d_m(\cdot - t)\rangle_{L^2(\mathbb{T})}\right| &= \frac{1}{(m+1)^2}\left|\int_{\mathbb{T}}\left(\sum_{k=0}^m \mathrm{e}^{2\pi ikt'}\right)\left(\sum_{\ell=0}^m \mathrm{e}^{-2\pi i\ell(t'-t)}\right) dt'\right| \\
&= \frac{|d_m(t)|}{m+1}(m+1)^2\,\|d_m\|^2_{L^2(\mathbb{T})} = \frac{|d_m(t)|}{m+1} \leq \frac{1}{m+1}\cdot\frac{1}{2}\cdot\frac{1}{(m+1)\,|t|_{\mathbb{T}}}
\end{aligned}$$

and with $z = \mathrm{e}^{2\pi it}$ and Parseval's identity also

$$\begin{aligned}
\left|\left\langle d_m^2, d_m^2(\cdot - t)\right\rangle_{L^2(\mathbb{T})}\right| &= \frac{1}{(m+1)^4}\int_{\mathbb{T}}\left(\frac{\sin(\pi(m+1)t')}{\sin(\pi t')}\cdot\frac{\sin(\pi(m+1)(t'-t))}{\sin(\pi(t'-t))}\right)^2 dt' \\
&= \frac{1}{(m+1)^4}\left|\sum_{k=-m}^m (m+1-|k|)^2 z^k\right|
\end{aligned}$$

$$= \frac{\left|(z+1)z\left(z^{m+1}-z^{-m-1}\right)+4(m+1)z(1-z)\right|}{(m+1)^4\left|z-1\right|^3}$$
$$\leq \frac{1}{m+1}\cdot\frac{1}{2}\cdot\frac{1}{(m+1)^2\left|t\right|_{\mathbb{T}}^2}.$$

Now, let t be the coordinate with $|t|_{\mathbb{T}} = |\boldsymbol{t}|_{\mathbb{T}^d}$, then the Cauchy–Schwarz inequality, iii), and the above yield (noting that $\mathrm{e}^{-2\pi i m t'} d_m^2(t') \geq 0$ and omitting the penultimate line if $\beta = 1$)

$$\begin{aligned}
\left|\left\langle d_m^\beta, d_m^\beta(\cdot - \boldsymbol{t})\right\rangle_{L^2(\mathbb{T}^d)}\right| &\leq \left|\int_{\mathbb{T}^d} d_m^\beta(\boldsymbol{t}')\overline{d_m^\beta(\boldsymbol{t}'-\boldsymbol{t})}\,d\boldsymbol{t}'\right| \\
&\leq \left\|d_m^\beta\right\|_{L^2(\mathbb{T})}^{2(d-1)}\left|\int_{\mathbb{T}} d_m^\beta(t')\overline{d_m^\beta(t'-t)}\,dt'\right| \\
&\leq \left\|d_m^\beta\right\|_{L^2(\mathbb{T})}^{2(d-1)}\left\|d_m d_m(\cdot - t)\right\|_{\mathcal{C}(\mathbb{T})}^{\beta-2}\left|\left\langle d_m^2, d_m^2(\cdot - t)\right\rangle_{L^2(\mathbb{T})}\right| \\
&\leq \frac{1}{2(m+1)^d\beta^{(d-1)/2}}\cdot\frac{1}{(m+1)^\beta\left|t\right|_{\mathbb{T}}^\beta}.
\end{aligned}$$

Finally we prove v). Since the modified Dirichlet kernel is a trigonometric polynomial, its coefficients coincide with its Fourier coefficients. Thus, its Fourier coefficients are non-negative. Increasing the power iteratively is according to Lemma 2.3.7 a convolution of its coefficients. In each convolution non-negative values are summed and hence, each power of the modified Dirichlet kernel has non-negative Fourier coefficients inductively. After each convolution the coefficients are bounded by 1. To see that, we start with the univariate $(m+1)d_m(t)$. Its Fourier coefficients are given by

$$(m+1)\widehat{d_m}(k) = \begin{cases} 1, & k = 0, \ldots, m, \\ 0, & \text{else.} \end{cases}$$

Now, increasing the power corresponds to convolving the present coefficient vector with the coefficient vector of d_m^β, i.e., $\frac{1}{m+1}(\ldots, 0, 1, \ldots, 1, 0, \ldots)^\top$. Assuming that for arbitrary $\beta > 1$ each coefficient of $(m+1)d_m^{\beta-1}$ is bounded by one, we obtain

$$(m+1)\widehat{((d_m^{\beta-1})} * \widehat{d_m})(\nu) = \sum_{\alpha\in\mathbb{Z}}\widehat{(d_m^{\beta-1})}(\nu-\alpha)\widehat{d_m}(\alpha) = \frac{1}{m+1}\sum_{\alpha=0}^{m}\widehat{(d_m^{\beta-1})}(\nu-\alpha) \leq 1$$

and the assertion follows by induction. The result for the power of multivariate modified Dirichlet kernels directly follows, since the Fourier coefficients of the tensor product are given by the tensor product of the Fourier coefficients. Concretely, we have with $\boldsymbol{t} = (t_1, \ldots, t_d)^\top$ for each $\boldsymbol{\nu} \in \mathbb{Z}^d$

$$\begin{aligned}
(m+1)^d\widehat{(d_m^\beta)}(\boldsymbol{\nu}) &= (m+1)^d\int_{\mathbb{T}^d} d_m^\beta(\boldsymbol{t})\mathrm{e}^{2\pi i\boldsymbol{\nu}^*\boldsymbol{t}}\,d\boldsymbol{t} = (m+1)^d\int_{\mathbb{T}^d}\prod_{\ell=1}^{d} d_m^\beta(t_\ell)\mathrm{e}^{2\pi i\nu_\ell t_\ell}\,dt_\ell \\
&= \prod_{\ell=1}^{d}(m+1)\int_{\mathbb{T}^d} d_m^\beta(t_\ell)\mathrm{e}^{2\pi i\nu_\ell t_\ell}\,dt_\ell = \prod_{\ell=1}^{d}(m+1)\widehat{(d_m^\beta)}(\nu_\ell),
\end{aligned}$$

which is smaller than one because each factor is, as shown above. □

Chapter 3

The condition number of Vandermonde matrices

For $M, N \in \mathbb{N}_+$ and nodes $z_1, \dots, z_M \in \mathbb{C}$ the (rectangular) Vandermonde matrix of degree $N-1$ is given by

$$\boldsymbol{A} := \boldsymbol{A}_N(z_1, \dots, z_M) := \left(z_j^{k-1}\right)_{j,k=1}^{M,N} = \begin{pmatrix} 1 & z_1^1 & z_1^2 & \cdots & z_1^{N-1} \\ 1 & z_2 & z_2^2 & \cdots & z_2^{N-1} \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & z_M & z_M^2 & \cdots & z_M^{N-1} \end{pmatrix} \in \mathbb{C}^{M \times N}. \tag{3.0.1}$$

A classical example in which square Vandermonde matrices appear, i.e. $M = N$, is the solution to the polynomial interpolation problem. Given a predefined set of nodes $\{z_1, \dots, z_M\} \in \mathbb{C}$ and corresponding function values $f(z_j) \in \mathbb{C}, j = 1, \dots, M$, the task is to find a polynomial p of degree less than M such that it interpolates the given points of the graph, i.e. $p(z_j) = f(z_j)$ for all j. Since the nodes are known, this task reduces to solving the linear system of equations

$$\boldsymbol{A}\boldsymbol{c} = \boldsymbol{f} \tag{3.0.2}$$

with square Vandermonde matrix $\boldsymbol{A} \in \mathbb{C}^{M \times M}$, right hand side $\boldsymbol{f} = \big(f(z_1), \dots, f(z_M)\big)^\top \in \mathbb{C}^M$ and the coefficient vector $\boldsymbol{c} \in \mathbb{C}^M$ of the interpolating polynomial. It is well-known ([55, p. 37]) that the determinant of the square Vandermonde matrix is given by

$$\det(\boldsymbol{A}) := \prod_{1 \le \ell < j \le M} (z_j - z_\ell). \tag{3.0.3}$$

Therefore, the matrix is invertible and hence the above system of equations has a unique solution if and only if all nodes are distinct. In that case the inverse is also known well ([55, p. 37], [62], [78]) and is given by

$$\boldsymbol{A}^{-1} = \big(a_{jk}\big)_{j,k=1}^{M}, \quad a_{jk} = (-1)^{j-1} \frac{S_{M-j}(z_1, \dots, z_{k-1}, z_{k+1}, \dots, z_M)}{\prod_{\ell=1, \ell \neq k}^{M} (z_\ell - z_k)}, \quad j, k = 1, \dots, M, \tag{3.0.4}$$

and where $S_0 = 1$ and if $m \ge \ell > 0$, S_ℓ is the ℓ-th elementary symmetric function

$$S_\ell \colon \mathbb{C}^m \to \mathbb{C}, \quad S_\ell(x_1, \dots, x_m) = \sum_{1 \le j_1 < \cdots < j_\ell \le m} \prod_{k=1}^{\ell} x_{j_k}.$$

In Chapter 2 we have seen in the motivation for defining the condition number in Definition 2.1.26, that it plays an important role for determining how strongly errors on the right hand side of a linear system of equations, like that in (3.0.2), are amplified. If $M > N$, the system (3.0.2) becomes over-determined and we refer to the introduction of [71] and the manuscript [50] for more details. In contrast we are interested in the rectangular case where we have less nodes, $M < N$, corresponding to an under-determined linear system of equation in (3.0.2), since these kind of Vandermonde matrices appear in exponential sum reconstruction which is a model problem for a variety of applications. Moreover, their condition number is essential for analyzing the stability of solution algorithms for this problem. More details are given in Chapter 4.

We focus on studying the spectral condition number of Vandermonde matrices and start with recapitulating known results in that context. After dealing with square Vandermonde matrices, we address rectangular cases. The remaining part of this chapter is devoted to the rectangular case ($M < N$) with nodes lying on the unit circle. In that context, we start with collecting known bounds for the condition number of Vandermonde matrices with well-separated nodes on the unit circle. Afterwards, we address the geometric model of clustered node configurations and analyze node sets that have at most two nodes per cluster with a Schur-complement technique. For more general clustered node configurations, we go one step further and present our results for the multivariate case. This is followed by a survey of existing results from the literature for univariate clustered node configurations that are additionally compared to our result. Finally, we present condition number bounds for multivariate well-separated node sets.

3.1 Survey of related results

In the second half of the last century research on the condition number of Vandermonde matrices became important and we see that this is still a subject of high interest. We start with a non exhaustive survey over work that has done in this direction. We always assume that nodes are distinct, to guarantee that involved Vandermonde matrices have full rank.

3.1.1 Square Vandermonde matrices

In most of the research around the condition number of square Vandermonde matrices Walter Gautschi was involved. For detailed overviews of his work we refer to his own survey [46], the collection [47] and especially the comment by Nicholas J. Higham in Section 5.1 of the latter. Furthermore, we refer to the book chapter [53, Ch. 22].

We stay with the square case $M = N$ in this section. Gautschi mostly studied the ∞-condition number of $\boldsymbol{A}^\top$. We remark that $\|\boldsymbol{A}\|_1 = \left\|\boldsymbol{A}^\top\right\|_\infty$ and due to Lemma 2.1.8 bounds can be transferred to the spectral norm condition number by

$$\frac{1}{M}\operatorname{cond}_1(\boldsymbol{A}) \leq \operatorname{cond}(\boldsymbol{A}) \leq M \operatorname{cond}_1(\boldsymbol{A}).$$

In [43, Thm.1] and [45, Thm. 3.1] Gautschi proved for the inverse of a square Vandermonde matrix with complex nodes

$$\max_{1\leq j\leq M} \prod_{\ell\neq j} \frac{\max\{1,|z_\ell|\}}{|z_j - z_\ell|} \leq \left\|\boldsymbol{A}^{-1}\right\|_1 \leq \max_{1\leq j\leq M} \prod_{\ell\neq j} \frac{1+|z_j|}{|z_j - z_\ell|}, \tag{3.1.1}$$

where in the upper bound equality holds if the nodes satisfy $z_j = |z_j| \, \mathrm{e}^{i\omega}$ for all j with fixed ω, i.e. they lie on a ray emanating from the origin. These bounds are at most 2^M apart. Using

$$\|\boldsymbol{A}\|_1 = \max_{0 \le k < M} \sum_{j=1}^{M} |z_j|^k = \max \left\{ M, \sum_{j=1}^{M} |z_j|^{M-1} \right\}$$

this yields (cf. [6, p. 5])

$$\max \left\{ M, \sum_{j=1}^{M} |z_j|^{M-1} \right\} \max_{1 \le j \le M} \prod_{\ell \ne j} \frac{\max\{1, |z_\ell|\}}{|z_j - z_\ell|} \le \operatorname{cond}_1(\boldsymbol{A}) \le \max \left\{ M, \sum_{j=1}^{M} |z_j|^{M-1} \right\} \max_{1 \le j \le M} \prod_{\ell \ne j} \frac{1 + |z_j|}{|z_j - z_\ell|}.$$

Based on identities for $\operatorname{cond}_1(\boldsymbol{A})$ established in [44] the lower bounds

$$\operatorname{cond}_1(\boldsymbol{A}) \ge 2^{M-1},$$

for $M \ge 2$ and non-negative nodes on the real line, and

$$\operatorname{cond}_1(\boldsymbol{A}) \ge (M-1)2^{M-1},$$

for $M \ge 2$ and real nodes symmetric around the origin, are provided in [48].

The exponentially increasing condition number for real nodes is also confirmed in [109, Thm. 4.1], where it is proven that the spectral condition number for any distinct real nodes fulfills

$$\operatorname{cond}(\boldsymbol{A}) \ge \frac{1}{\sqrt{M}} 2^{M-2}$$

and if either $|z_j| \le 1$ or $|z_j| \ge 1$ for all $j = 1, \ldots, M$, then

$$\operatorname{cond}(\boldsymbol{A}) \ge 2^{M-2}.$$

For distinct real nodes the exponential growth is also proven differently in [101] by means of Chebyshev polynomials and in [29] by application of the inequality

$$\sum_{j=1}^{M} \frac{1}{\prod_{\ell=1, \ell \ne j}^{M} |z_j - z_\ell|} \ge 2^{M-2}, \quad \text{for all } |z_j| \le 1,$$

from [38, La. 1] in the lower bound of (3.1.1). Sharp bounds are established by Beckermann in [17, Thm. 4.1]. Providing also an upper bound, that differ only by a factor of $(M+1)^{3/2}$, shows that his bounds are quasi optimal. For distinct real nodes it is stated

$$\frac{\sqrt{2}}{\sqrt{M+1}} (1+\sqrt{2})^{M-1} \le \inf \operatorname{cond}(\boldsymbol{A}) \le \sqrt{2}(M+1)(1+\sqrt{2})^{M-1}$$

and if the nodes are on the positive real line $\mathbb{R}_{\ge 0}$, then

$$\frac{1}{2\sqrt{M+1}} C_M \le \inf \operatorname{cond}(\boldsymbol{A}) \le \frac{M+1}{2} C_M,$$

where $C_M := \left((1+\sqrt{2})^{2M} + (1+\sqrt{2})^{-2M}\right)$. The infima are taken over all Vandermonde matrices with nodes from the respective sets.

Vandermonde matrices are not exponentially ill-conditioned in general. The situation on the complex unit circle shows the opposite. An heuristic argument for the statement that Vandermonde matrices with nodes on the unit circle are well conditioned might be the following. All nodes have absolute value one and therefore, each entry of the Vandermonde matrix has absolute value one as well, no matter what degree the Vandermonde matrix has. Indeed, well-known example of a well-conditioned Vandermonde matrix is the discrete Fourier matrix $\boldsymbol{F}_M \in \mathbb{C}^{M\times M}$, with equispaced nodes $z_j = \mathrm{e}^{2\pi i \frac{j}{M}}, j = 0, \dots, M-1$, ($M$-th roots of unity) and entries

$$\left(\boldsymbol{F}_M\right)_{jk} = \mathrm{e}^{2\pi i \frac{kj}{M}}, \quad j,k = 0,\dots,M-1. \tag{3.1.2}$$

More precisely, it is perfectly conditioned, i.e. $\operatorname{cond}(\boldsymbol{F}) = 1$, since

$$\boldsymbol{F}_M \boldsymbol{F}_M^* = M\left(d_{M-1}(\tfrac{j-\ell}{M})\right)_{j,\ell=1}^{M} = \operatorname{diag}(M,\dots,M)$$

by Definition 2.4.10 and $\sin(\pi M \frac{k}{M}) = 0$ for all $k \in \mathbb{Z}$. The singular values of $\boldsymbol{F}_M$ are therefore all given by $\sqrt{M}$.

However, the above argument is inadequate and the good situation does not hold for all Vandermonde matrices with nodes on the unit circle. Particularly, if nodes are close to each other, the condition number must increase. This can be seen by combining the continuity of singular values with respect to the matrix entries and the singularity of the matrix if two nodes are equal. Furthermore, in [18] the authors showed that a square Vandermonde matrix with nodes on the unit circle is perfectly conditioned, if and only if the nodes are uniformly spread over the circle. Their main point is quite simple and can be compared to the following argumentation. Let the nodes be parameterized by $t_j \in \mathbb{T}$ such that $z_j = \mathrm{e}^{2\pi i t_j}$ for $j = 1,\dots,M$. Then as before $\boldsymbol{A}\boldsymbol{A}^* = M\left(d_{M-1}(t_j - t_\ell\right)_{j,\ell=1}^{M}$. The nodes are assumed to be distinct, thus, $\boldsymbol{A}$ has full rank, and using Lemma 2.1.27, the statement is true if $\boldsymbol{A}\boldsymbol{A}^*$ is diagonal with constant diagonal entries. Each diagonal entry is M since $d_{M-1}(0) = 1$. For the off-diagonal entries we have

$$\left(\boldsymbol{A}\boldsymbol{A}^*\right)_{j,\ell} = \mathrm{e}^{\pi i (N-1)(t_j - t_\ell)} \frac{\sin(\pi M(t_j - t_\ell))}{\sin(\pi(t_j - t_\ell))}.$$

The denominator and the prefactor are non zero and hence, the entry is zero if and only if $\sin(\pi M(t_j - t_\ell)) = 0$. This is the case if and only if $M(t_j - t_\ell) \in \mathbb{Z}$. We have $t_j - t_\ell < 1$ leading to the only possibility $t_j - t_\ell = \frac{k}{M}, k \in \mathbb{Z}$, and in particular $\min_{j\neq\ell} |t_j - t_\ell| \geq \frac{1}{M}$. Combined with the fact, that we have M distinct nodes on the unit circle, they must be equispaced. On the other hand for M equispaced nodes we have $|t_j - t_\ell| = \frac{1}{M}, j \neq \ell$, clearly satisfying $M(t_j - t_\ell) \in \mathbb{Z}$.

Even if the nodes are almost equispaced this is no guarantee for the Vandermonde matrix to be well conditioned. There are two node sequences on the unit circle, that got special attention in the past, see [46, Sec. IV], underlining this. One is given by means of a Van der Corput sequence being an example for nodes that are not equispaced and leading to a mildly growing condition number of the corresponding square Vandermonde matrix with respect to the number of nodes. The other is called quasi-cyclic sequence and has well distributed nodes, consisting only of 2^k-th and 2^{k-1}-th roots of unity for $2^{k-1} < M \leq 2^k$, though it leads to

Vandermonde matrices with exponentially growing condition number. The Van der Corput sequence $(c_n)_{n\in\mathbb{N}}$ with respect to the binary representation, is constructed as follows. Let $n \in \mathbb{N}$. We set $c_0 = 1$ and for $n > 0$ given the binary representation

$$n = \sum_{j=0}^{\ell(n)} b_j 2^j$$

with $b_j \in \{0, 1\}$ and uniquely determined $\ell(n) \in \mathbb{N}$, the n-th number c_n in the sequence is given by

$$c_n := \sum_{j=0}^{\ell(n)} b_j \frac{1}{2^{j+1}}.$$

In [27] the spectral condition number of the square Vandermonde matrix $\boldsymbol{A}_{\mathrm{vdc}} \in \mathbb{C}^{M\times M}$ with nodes $z_j = \mathrm{e}^{2\pi i c_{j-1}}, j = 1, \ldots, M$, predetermined by the Van der Corput sequence (see Figure 3.1.1, left) is studied. The authors provide explicit representations of the eigenvalues of $\boldsymbol{A}_{\mathrm{vdc}}\boldsymbol{A}_{\mathrm{vdc}}^*$ and deduce

$$\mathrm{cond}(\boldsymbol{A}_{\mathrm{vdc}}) \leq \sqrt{2M}.$$

For constructing the quasi-cyclic sequence we observe that for $k \in \mathbb{N}$ the 2^k-th roots of unity are increasingly contained into each other, i.e. the 2^k-th roots of unity are contained in the 2^{k+1}-st roots of unity. The quasi-cyclic sequence $(w_n)_{n\in\mathbb{N}_+}$ is then iteratively given by going around the unit circle in positive direction and for $2^k < n \leq 2^{k+1}$ setting w_n as the next free 2^{k+1}-st root of unity, see Figure 3.1.1 (right). Formally, if $2^k < n \leq 2^{k+1}$, the sequence is given by

$$w_1 = 1, \quad w_n = \mathrm{e}^{2\pi i \frac{2(n-2^k)-1}{2^{k+1}}}.$$

Taking the corresponding square Vandermonde matrix $\boldsymbol{A}_{\mathrm{qc}} \in \mathbb{C}^{M\times M}$ into account, it is clear that for M being a power of two, we have a full set of roots of unity and therefore, $\mathrm{cond}(\boldsymbol{A}_{\mathrm{qc}}) = 1$. But in [46] it is numerically shown that, for values of M in between, the condition number becomes significantly larger for increasing M. Quite recently in [85, Thm. 5.1] this behavior is confirmed theoretically. It is shown, if $M = 3 \cdot 2^k$ for some $k \in \mathbb{N}$, then

$$\mathrm{cond}(\boldsymbol{A}_{\mathrm{qc}}) \geq \sqrt{M} 4^{k-1}.$$

Note that for $M = 3 \cdot 2^k$ all 2^{k+1}-st roots of unity plus the missing 2^{k+2}-nd roots in the upper unit circle are in the considered node set.

3.1.2 Rectangular Vandermonde matrices

We state some recently developed results for rectangular Vandermonde matrices in this section.

Contiguous submatrices of the discrete Fourier Matrix

A few month ago, Barnett investigated rectangular, contiguous submatrices of the discrete Fourier matrix with arbitrary shape in [8]. These matrices can be regarded as rectangular Vandermonde matrices as given in (3.0.1) and we therefore denote them also by $\boldsymbol{A} \in \mathbb{C}^{M\times N}$. Upper bounds for the smallest, and lower bounds for the largest singular values are given,

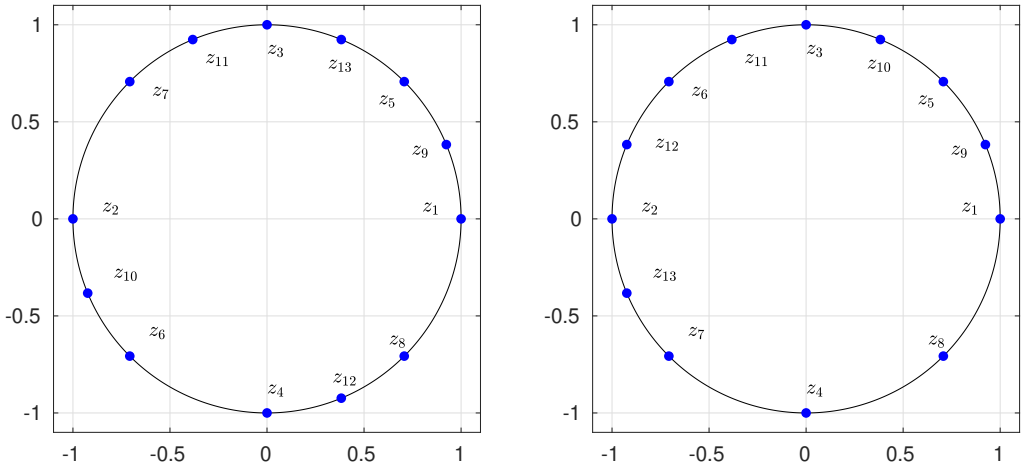


Figure 3.1.1: left: first 13 nodes given by the Van der Corput sequence; right: first 13 nodes of the quasi-cyclic sequence.

proving that the condition number of all these matrices grow exponentially. We note that these results are symmetric in M and N since $\operatorname{cond}(\boldsymbol{A}) = \operatorname{cond}(\boldsymbol{A}^*)$ and the entries $\mathrm{e}^{2\pi i \frac{jk}{\gamma}}$ of the discrete Fourier matrix $\boldsymbol{F}_\gamma, \gamma \in \mathbb{N}_+$, allow to interpret both $\frac{j}{\gamma}$ and $\frac{k}{\gamma}$ as parameter for the nodes z_j. While the lower bound on the largest singular value $\sigma_{\min}(\boldsymbol{A}) \geq \sqrt{N}$ is readily given by rearranging

$$\sqrt{MN} = \|\boldsymbol{A}\|_{\mathrm{F}} \leq \sqrt{N}\,\|\boldsymbol{A}\|\,, \tag{3.1.3}$$

the upper bound on the smallest singular value is established by using its variation characterization, see Theorem 2.1.15. If $M \leq N$ (due to symmetry without loss of generality), special test vectors $\boldsymbol{v} \in \mathbb{C}^M$ are applied. Then we have $\sigma_{\min}(\boldsymbol{A}) \leq \|\boldsymbol{A}^*\boldsymbol{v}\|$ and instead of further bounding the right hand side, an embedding of the matrix vector product $\boldsymbol{A}\boldsymbol{v}$ into the product $\boldsymbol{F}_\gamma \boldsymbol{f}$ is applied. Therein $\boldsymbol{f} \in \mathbb{C}^\gamma$ is the discrete Fourier transform of a special function.

Theorem 3.1.1 (Contiguous submatrices of the DFT matrix 1, [8, Thm. 1])**.**
Let $N, M, \gamma \in \mathbb{N}_+$ and $2 < M \leq N < \gamma - 2$ then for any contiguous submatrix (also taken periodically) $\boldsymbol{A} \in \mathbb{C}^{M\times N}$ of the discrete Fourier matrix $\boldsymbol{F}_\gamma$ given in (3.1.2) *it holds*

$$\operatorname{cond}(\boldsymbol{A}) \geq \frac{\sqrt{N}\left(1-\frac{\widetilde{N}}{\gamma}\right)^{\frac{1}{4}}}{6\widetilde{M}^{\frac{1}{4}}\sqrt{\gamma}}\mathrm{e}^{\frac{\pi}{4}\left(1-\frac{\widetilde{N}}{\gamma}\right)\widetilde{M}},$$

where $\widetilde{M}$ is the largest even integer smaller than M, and $\widetilde{N}$ is the smallest integer of the same parity as γ larger than N.

A second theorem will be useful for us later on because we use it to state lower bounds for the condition number in the case of clustered nodes on the unit circle.

Theorem 3.1.2 (Contiguous submatrices of the DFT matrix 2, cf. [8, Thm. 3])**.**
Let $N, M, \gamma \in \mathbb{N}_+$ and $1 < M \leq N < \frac{4\gamma}{\mathrm{e}\pi} + 1$ then for any contiguous submatrix (also taken

periodically) $\boldsymbol{A} \in \mathbb{C}^{M\times N}$ *of the discrete Fourier matrix* $\boldsymbol{F}_\gamma$ *given in* (3.1.2) *it holds*

$$\operatorname{cond}(\boldsymbol{A}) \geq \frac{1-\left(\frac{\mathrm{e}\pi(N-1)}{4\gamma}\right)}{2\sqrt{M}} \left(\frac{4\gamma}{\mathrm{e}\pi(N-1)}\right)^{M-1}.$$

Barnett also states a third theorem which improves the exponential rate in Theorem 3.1.1 from $\frac{\pi}{4}$ to $\frac{\pi}{2}$ but with different prefactors.

Nodes on the unit disc

If all nodes lie on the complex unit disc, an upper bound on the condition number is established in [7] based on the application of a further generalized Hilbert's inequality. More precisely, a lower bound on the smallest singular value and an upper bound on the largest singular value is given, such that an upper bound for the condition number follows directly by the identity $\operatorname{cond}(\boldsymbol{A}) = \sigma_{\max}(\boldsymbol{A})/\,\sigma_{\min}(\boldsymbol{A})$ from Definition 2.1.26. From these bounds also bounds for the case of nodes on the complex unit circle are derived, that we come back to in the next section.

Theorem 3.1.3 (Upper bound on the condition number, [7, Thm.5])**.**
Let the nodes be given by $z_j = |z_j|\,\mathrm{e}^{2\pi i t_j} \in \mathbb{C}$ *with* $0 \leq 1$ *and* $t_j \in [0,1)$ *with distance*

$$q_j := \min_{\substack{1\leq \ell\leq M \\ \ell\neq j}} \min_{k\in\mathbb{Z}} |t_j - t_\ell + k| > 0$$

for all $j = 1,\ldots,M$*. Then we have for the extremal singular values of the Vandermonde matrix* $\boldsymbol{A} \in \mathbb{C}^{M\times N}$ *with* $M \leq N$

$$\begin{aligned}
\sigma_{\min}(\boldsymbol{A})^2 &\geq \min_{1\leq j\leq M} \left\{ \frac{1}{|z_j|} \left[\varphi_N(|z_j|) - \frac{42}{\pi q_j} \left(1 + |z_j|^{2N}\right) \right] \right\} \\
\sigma_{\max}(\boldsymbol{A})^2 &\leq \min\left\{u(N), u(N-1)\right\},
\end{aligned}$$

where

$$u(N) := \max_{1\leq j\leq M} \left\{ \frac{1}{|z_j|} \left[\varphi_N(|z_j|) + \frac{42}{\pi q_j} \left(1 + |z_j|^{2N}\right) \right] \right\}$$

and

$$\varphi_N(|z_j|) := \begin{cases} \frac{|z_j|^{2N}-1}{2\log|z_j|}, & |z_j| < 1, \\ N, & |z_j| = 1. \end{cases}$$

Moreover, if $|z_1| = \cdots = |z_M| = r < 1$ *and* $q := \min_{1\leq j\leq M} q_j$ *then*

$$\sigma_{\min}(\boldsymbol{A})^2 \geq \frac{1 - r^{2(N+1/2-1/q)}}{q(r^{-2/q}-1)r^2} \tag{3.1.4}$$

$$\sigma_{\max}(\boldsymbol{A})^2 \leq \frac{r^{-2/q}\left(1 - r^{2(N+1/2-1/q)}\right)}{q(r^{-2/q}-1)}. \tag{3.1.5}$$

There is also the analytic result from [15] that provides an upper and a lower bound for the condition number of Vandermonde matrices with nodes on the unit disc. Besides the complicated structure of the bounds it has two disadvantages compared to the above result from [7]. It is of implicit nature in the sense that the solution $\boldsymbol{c} \in \mathbb{C}^N$ to the Vandermonde

system $\boldsymbol{A}\boldsymbol{c} = (z_1^N, \dots, z_M^N)^\top$ has to be known. Secondly, the condition number bounds are directly given and not established by single bounds on the largest and smallest singular value. For the special case with M nodes being on a fixed circle around the origin, i.e. $|z_1| = \cdots = |z_M| = r < 1$, and $1 - r^2 \leq (\min_{j \neq \ell} |z_j - z_\ell|)^2$ in [15, Cor. 10] the asymptotic upper bound

$$\lim_{N \to \infty} \operatorname{cond}(\boldsymbol{A}_N) \leq M 2^{\frac{M-1}{2} - M + 2}$$

is derived. A detailed discussion and a comparison to the results from [7] are given in [7, Sec. 5.3].

3.1.3 Well-separated nodes on the unit circle

From now on we only deal with rectangular Vandermonde matrices on the unit circle. Let $N, M \in \mathbb{N}_+$, $M \leq N$ and distinct nodes $z_1, \dots, z_M \in \{z \in \mathbb{C} \mid |z| = 1\}$, where the $z_j = \mathrm{e}^{2\pi i t_j}, j = 1, \dots, M$, are parameterized by $t_1, \dots, t_M \in \mathbb{T}$. Then the Vandermonde matrix of degree $N - 1$ is given by

$$\boldsymbol{A} = \left(z_j^{k-1}\right)_{j,k=1}^{M,N} = \left(\mathrm{e}^{2\pi i (k-1) t_j}\right)_{j,k=1}^{M,N} \in \mathbb{C}^{M \times N} \tag{3.1.6}$$

as before. Since the analysis is relaying mostly on the parameter $t_j \in \mathbb{T}$ and they are connected to the z_j by a simple bijection, we also call them nodes and collect them in the set $\Omega := \{t_1, \dots, t_M\}$. We have seen in the beginning of this chapter that $\boldsymbol{A}$ has full rank if and only if the nodes are distinct. Furthermore, the singular values of a matrix are continuously depending on its entries (as the eigenvalues are, see [55, Thm. 2.4.9.2], combined with Lemma 2.1.14) so that the condition number of the Vandermonde matrix is expected to grow if nodes become close. A notion of distance between the nodes is therefore meaningful. In particular, since the nodes are on the complex unit circle it offers to use an adapted distance rather than the Euclidean distance between complex numbers. The following is basically the (normalized) arc length between points on the unit circle.

Definition 3.1.4 (Wrap-around distance)**.**
The wrap-around distance *between two nodes $t, t' \in \mathbb{T}$ is defined by*

$$\left|t - t'\right|_{\mathbb{T}} := \min_{r \in \mathbb{Z}} \left|t - t' + r\right|.$$

The minimal separation distance *q of a node set $\Omega \subset \mathbb{T}$ is given by*

$$q := \min_{t, t' \in \Omega,\, t \neq t'} \left|t - t'\right|_{\mathbb{T}}.$$

Definition 3.1.5 (Well-separated node set)**.**
Let $M, N \in \mathbb{N}_+$ and $\boldsymbol{A} \in \mathbb{C}^{M \times N}$ be a Vandermonde matrix with nodes $\Omega \subset \mathbb{T}$. The node set Ω is called well-separated *if the minimal separation distance q fulfills*

$$q > \frac{1}{N}.$$

Bazán stated in [15, (3.18)] the following upper bound on the condition number by applying Gerschgorin's circle theorem (Lemma 2.1.22) to bound the extremal eigenvalues of the

Hermitian kernel matrix $\boldsymbol{A}\boldsymbol{A}^* = \big(Nd_{N-1}(t_j - t_\ell)\big)_{j,\ell=1,\dots,M}$. Let $\eta := \min_{1\le j<k\le M} |z_j - z_k|$ denote the minimal Euclidean distance between the nodes, then

$$\operatorname{cond}(\boldsymbol{A})^2 \le \frac{\eta N + 2M - 2}{\eta N - 2M + 2} \quad \text{if} \quad \eta N > 2(M-1).$$

He used the identity $2\sin((\theta-\theta')/2) = |z - z'|$ for two nodes $z = \mathrm{e}^{i\theta}, z' = \mathrm{e}^{i\theta'}$ on the unit circle to bound the absolute value of the off-diagonal entries in $\boldsymbol{A}\boldsymbol{A}^*$. Using the inequality $\sin(x) \ge \frac{2}{\pi}x$, for $0 \le x \le \frac{\pi}{2}$, instead, we can rewrite the bound in terms of the wrap-around distance to

$$\operatorname{cond}(\boldsymbol{A})^2 \le \frac{2qN + M - 1}{2qN - M + 1} \quad \text{if} \quad qN > \frac{M-1}{2}. \tag{3.1.7}$$

The quite strong dependence on the number of nodes M is due to bounding each off-diagonal entry uniformly, regardless of the fact, that for a fixed node its distance to the other nodes increase from node to node on both sides of the interval. Potts and Tasche bound the largest singular value from above by means of the Gerschgorin's circle theorem in [91, Cor. 4.3]. It readily provides also a lower bound for the smallest singular value as seen above. Additionally, they took into account the increasing distances from one node to the others in both directions on the interval. Following their proof, we obtain

$$\operatorname{cond}(\boldsymbol{A})^2 \le \frac{qN + \log(\lfloor \frac{M}{2} \rfloor) + 1}{qN - \log(\lfloor \frac{M}{2} \rfloor) - 1} \quad \text{if} \quad qN > \log\left(\left\lfloor \frac{M}{2} \right\rfloor\right) + 1, \tag{3.1.8}$$

where $\lfloor x \rfloor$ is the largest integer smaller or equal $x \in \mathbb{R}$.

Ferreira, [41], and Liao and Fannjiang, [76, Thm. 2] provide upper bounds on the condition number by bounding the extremal singular values as well. Having a closer look reveals that both results rely on the following lemma that formalizes the technique used in [41].

Lemma 3.1.6 (cf. [41, p. 218])**.**
Let $\chi_{[0,N-1]}$ be the characteristic function on the interval $[0, N-1]$ and $s, g \in \mathcal{C}(\mathbb{T})$ be trigonometric polynomials with real Fourier coefficients satisfying

$$\widehat{s}(k) \le \chi_{[0,N-1]}(k) \le \widehat{g}(k)$$

for all $k \in \mathbb{Z}$. We denote their kernel matrices with respect to the nodes Ω by

$$\boldsymbol{S} := \big(s(t_j - t_\ell)\big)_{j,\ell=1}^{M} \quad \text{and} \quad \boldsymbol{G} := \big(g(t_j - t_\ell)\big)_{j,\ell=1}^{M}.$$

Then we have

$$\begin{aligned} \lambda_{\min}(\boldsymbol{S}) \le\ & \lambda_{\min}(\boldsymbol{A}\boldsymbol{A}^*) = \sigma_{\min}(\boldsymbol{A})^2 \le \lambda_{\min}(\boldsymbol{G}) \\ \lambda_{\max}(\boldsymbol{S}) \le\ & \lambda_{\max}(\boldsymbol{A}\boldsymbol{A}^*) = \sigma_{\max}(\boldsymbol{A})^2 \le \lambda_{\max}(\boldsymbol{G}). \end{aligned}$$

Proof. We define the formal, infinite extended Vandermonde matrix by

$$\boldsymbol{A}_\infty := \big(z_j^k\big)_{\substack{j=1,\dots,M \\ k\in\mathbb{Z}}}.$$

Furthermore, we define similarly the formal, infinite diagonal matrices $\boldsymbol{C}_s := \operatorname{diag}(\widehat{s}_k)_{k\in\mathbb{Z}}$ and $\boldsymbol{C}_g := \operatorname{diag}(\widehat{g}_k)_{k\in\mathbb{Z}}$. A simple calculation shows that $\boldsymbol{S} = \boldsymbol{A}_\infty \boldsymbol{C}_s \boldsymbol{A}_\infty^*$ and $\boldsymbol{G} = \boldsymbol{A}_\infty \boldsymbol{C}_g \boldsymbol{A}_\infty^*$.

Furthermore, both matrices are Hermitian as the diagonal matrices are real. Before we use Theorem 2.1.4 for the extremal eigenvalues of $\boldsymbol{A}\boldsymbol{A}^*$, we look at the Rayleigh quotient. For any vector $\boldsymbol{v} \in \mathbb{C}^M$ we have

$$\begin{aligned}\boldsymbol{v}^*\boldsymbol{A}\boldsymbol{A}^*\boldsymbol{v} = \|\boldsymbol{A}^*\boldsymbol{v}\|^2 &= \sum_{k=0}^{N-1} |(\boldsymbol{A}^*\boldsymbol{v})_k|^2 = \sum_{k\in\mathbb{Z}} \chi_{[0,N-1]}(k)\, |(\boldsymbol{A}_\infty^*\boldsymbol{v})_k|^2 \\ &\geq \sum_{k\in\mathbb{Z}} \widehat{s}(k)\, |(\boldsymbol{A}_\infty^*\boldsymbol{v})_k|^2 = \sum_{k\in\mathbb{Z}} \widehat{s}(k) \overline{(\boldsymbol{A}_\infty^*\boldsymbol{v})_k}\, (\boldsymbol{A}_\infty^*\boldsymbol{v})_k \\ &= \sum_{k\in\mathbb{Z}} \overline{(\boldsymbol{A}_\infty^*\boldsymbol{v})_k}\, (\boldsymbol{C}_s\boldsymbol{A}_\infty^*\boldsymbol{v})_k = \boldsymbol{v}^*\boldsymbol{A}_\infty\boldsymbol{C}_s\boldsymbol{A}_\infty^*\boldsymbol{v} = \boldsymbol{v}^*\boldsymbol{S}\boldsymbol{v}.\end{aligned} \tag{3.1.9}$$

Taking now minimum over all unit norm vectors $\boldsymbol{v} \in \mathbb{C}^M$, we obtain by means of Theorem 2.1.4 and Lemma 2.1.14 ii)

$$\sigma_{\min}(\boldsymbol{A})^2 = \lambda_{\min}(\boldsymbol{A}\boldsymbol{A}^*) \geq \lambda_{\min}(\boldsymbol{S}) \quad \text{and} \quad \sigma_{\max}(\boldsymbol{A})^2 = \lambda_{\max}(\boldsymbol{A}\boldsymbol{A}^*) \geq \lambda_{\max}(\boldsymbol{S}).$$

Analogously, we can use the Fourier coefficients $\widehat{g}$ in (3.1.9) leading to $\boldsymbol{v}^*\boldsymbol{A}\boldsymbol{A}^*\boldsymbol{v} \leq \boldsymbol{v}^*\boldsymbol{G}\boldsymbol{v}$ and finally yields the upper bounds. □

Once Lemma 3.1.6 is applied with appropriate trigonometric polynomials, hopefully good kernel matrices are obtained for which Gerschgorin's circle theorem leads to better bounds than applying it to the kernel matrix $\boldsymbol{A}\boldsymbol{A}^*$ directly as already seen in (3.1.8).

In [41] trigonometric polynomials with trapezoidal Fourier coefficients are used to upper and lower bound the characteristic function $\chi_{[0,N-1]}$. Together with Gerschgorin's circle theorem it yields (combining bounds on the extremal eigenvalues and reverting the unnecessary assumption that nodes are on a grid) the following. Let the maximal wrap-around distance be $w := \max_{1\leq j<j'\leq 1} |t_j - t_{j'}|_{\mathbb{T}}$ and for $x \in \mathbb{R}$ we denote $[x]$ the closest integer to $x \in \mathbb{R}$ (taking the larger if not unique). Then we have

$$\operatorname{cond}(\boldsymbol{A})^2 \leq \frac{N-1+[c]+\frac{c^2}{[c]^2}}{N+1-[c]-\frac{c^2}{[c]^2}} \quad \text{if} \quad N > [c] + \frac{c^2}{[c]^2} - 1,$$

where $c := \frac{\pi w}{\sqrt{3}\sin(\pi w)q}$. According to [7, p. 12], this bound can be further simplified to

$$\operatorname{cond}(\boldsymbol{A})^2 \leq \frac{3qN+7}{3qN-7} \quad \text{if} \quad qN > \frac{7}{3}. \tag{3.1.10}$$

Denoting by $\lceil x \rceil$ the smallest integer larger than $x \in \mathbb{R}$ the bound established in [76] can be written as (see also [7, (9)])

$$\operatorname{cond}(\boldsymbol{A})^2 \leq \frac{8\sqrt{2}\,\lceil (N-1)/2 \rceil + \frac{1}{\sqrt{2}\lceil (N-1)/2 \rceil q^2} + 3\sqrt{2}\pi}{2(N-1) - \frac{2}{(N-1)q^2} - 4\pi} \quad \text{if} \quad q(N-1) > \left(1 - \frac{2\pi}{N-1}\right)^{-\frac{1}{2}}$$

and $N \geq 7$. The calculations in there yield bounds of the form

$$a\,\|\boldsymbol{v}\|^2 \leq \|\boldsymbol{A}^*\boldsymbol{v}\| \leq b\,\|\boldsymbol{v}\| \tag{3.1.11}$$

for arbitrary $\boldsymbol{v} \in \mathbb{C}^M$ on a direct way. These bounds are also known as discrete Ingham inequalities [57, 84, 63, 64] and together with Theorem 2.1.15 the terms a, b directly give

bounds for the extremal singular values of $\boldsymbol{A}$. Having Lemma 3.1.6 at hand enables us to simplify the view on their technique. Basically their calculation can be seen as applying Lemma 3.1.6 with $\widehat{g}(k) := \cos\left(\pi\left(\frac{k}{N-1} - \frac{1}{2}\right)\right)$ and $\widehat{s}(k) := \left(\cos\left(\frac{\pi}{4}\right)\right)^{-1}\cos\left(\pi\left(\frac{k}{2(N-1)} - \frac{1}{4}\right)\right)$ and eventually using Gerschgorin's circle theorem on the arising kernel matrices.

Finally, we come to the state of the art upper bound on the condition number of Vandermonde matrices with well-separated nodes on the unit circle. It follows a similar approach as seen in the proof of Lemma 3.1.6 but uses in some sense optimal functions in $\mathcal{C}(\mathbb{T})$ that majorize and minorize the characteristic function $\chi_{[0,N-1]}$.

Theorem 3.1.7 (Upper bound on the condition number, [80, 7, 34])**.**
Let $\boldsymbol{A}$ be a Vandermonde matrix as in (3.1.6) *with $N \geq M$ and q-separated nodes. Then we have*

$$N + 1 - \frac{1}{q} \leq \sigma_{\min}(\boldsymbol{A})^2 \leq N \leq \sigma_{\max}(\boldsymbol{A})^2 \leq N - 1 + \frac{1}{q}.$$

In particular, if the nodes are well-separated, it holds

$$\operatorname{cond}(\boldsymbol{A})^2 \leq \frac{qN+1}{qN-1} = 1 + \frac{2}{qN-1}.$$

Proof. The inner inequalities follow from the variational characterization, Theorem 2.1.15, of the extremal singular values. Let $\mathbf{1}_M = (1, \ldots, 1)^\top \in \mathbb{C}^M$, then since $M \leq N$

$$\sigma_{\min}(\boldsymbol{A})^2 = \min_{\boldsymbol{v}\in\mathbb{C}^M} \|\boldsymbol{A}^*\boldsymbol{v}\|^2 \leq \|\boldsymbol{A}^*\mathbf{1}_M\|^2 = \sum_{k=1}^{N}\left|z_j^k\right|^2 = N$$

and analogously $\sigma_{\max}(\boldsymbol{A}) \geq N$. For the upper bound on $\sigma_{\max}(\boldsymbol{A})$ we use the Selberg majorant h from Lemma 2.4.3 for the characteristic function $\chi_{[0,N-1]}$ over the interval $[0, N-1]$ and with Fourier transform supported in $[-q, q]$. Therefore, using the Poisson summation formula from Lemma 2.3.9 and observing $\widehat{h}(|t_j - t_k|_{\mathbb{T}}) = 0$ for $j \neq k$, we obtain for any vector $\boldsymbol{v} = (v_1, \ldots, v_m)^\top \in \mathbb{C}^M$

$$\begin{aligned}
\|\boldsymbol{A}^*\boldsymbol{v}\|^2 &= \sum_{k=0}^{N-1}\left|(\boldsymbol{A}^*\boldsymbol{v})_k\right|^2 = \sum_{k=-\infty}^{\infty}\chi_{[0,N-1]}(k)\sum_{j=1}^{M}v_j\mathrm{e}^{2\pi ikt_j}\sum_{\ell=1}^{M}\overline{v_\ell}\mathrm{e}^{-2\pi ikt_\ell} \\
&\leq \sum_{k=-\infty}^{\infty}h(k)\sum_{j=1}^{M}v_j\mathrm{e}^{2\pi ikt_j}\sum_{\ell=1}^{M}\overline{v_\ell}\mathrm{e}^{-2\pi ikt_\ell} = \sum_{j=1}^{M}\sum_{\ell=1}^{M}v_j\overline{v_\ell}\sum_{k=-\infty}^{\infty}h(k)\mathrm{e}^{2\pi ik(t_j-t_\ell)} \qquad (3.1.12) \\
&= \sum_{j=1}^{M}\sum_{\ell=1}^{M}v_j\overline{v_\ell}\sum_{r=-\infty}^{\infty}\widehat{h}(t_j - t_\ell + r) = \widehat{h}(0)\sum_{j=1}^{M}|v_j|^2 = \left(N - 1 + \frac{1}{q}\right)\|\boldsymbol{v}\|^2.
\end{aligned}$$

Dividing by $\|\boldsymbol{v}\|^2$ and applying the variational characterization of $\sigma_{\max}$ from Theorem 2.1.15 yields the inequality since $\boldsymbol{v}$ was arbitrary. In order to obtain a lower bound on the smallest singular value we can proceed similarly. We use the Selberg minorant l for the characteristic

function $\chi_{[-1,N]}$. Then $l(k) \leq 0$ still for $k \notin \{0, \ldots, N-1\}$ and hence, we obtain

$$\begin{aligned}\|\boldsymbol{A}^*\boldsymbol{v}\|^2 &= \sum_{k=0}^{N-1} |(\boldsymbol{A}^*\boldsymbol{v})_k|^2 = \sum_{k=-\infty}^{\infty} \chi_{[0,N-1]}(k) \sum_{j=1}^{M} v_j \mathrm{e}^{2\pi i k t_j} \sum_{\ell=1}^{M} \overline{v_\ell} \mathrm{e}^{-2\pi i k t_\ell} \\ &\geq \sum_{k=-\infty}^{\infty} l(k) \sum_{j=1}^{M} v_j \mathrm{e}^{2\pi i k t_j} \sum_{\ell=1}^{M} \overline{v_\ell} \mathrm{e}^{-2\pi i k t_\ell} = \sum_{j=1}^{M} \sum_{\ell=1}^{M} v_j \overline{v_\ell} \sum_{k=-\infty}^{\infty} l(k) \mathrm{e}^{2\pi i k (t_j - t_\ell)} \\ &= \sum_{j=1}^{M} \sum_{\ell=1}^{M} v_j \overline{v_\ell} \sum_{r=-\infty}^{\infty} \widehat{l}(t_j - t_\ell + r) = \widehat{l}(0) \sum_{j=1}^{M} |v_j|^2 = \left(N + 1 - \frac{1}{q}\right) \|\boldsymbol{v}\|^2 .\end{aligned} \tag{3.1.13}$$

If $\boldsymbol{A}$ is well-separated then $q > \frac{1}{N}$ and $\sigma_{\min}(\boldsymbol{A}) > 0$ by using the just established bound. Finally, the condition number bound is obtained by

$$\operatorname{cond}(\boldsymbol{A}) = \frac{\sigma_{\max}(\boldsymbol{A})}{\sigma_{\min}(\boldsymbol{A})} \leq \frac{N + \frac{1}{q}}{N - \frac{1}{q}} = \frac{qN+1}{qN-1} = \frac{qN-1+2}{qN-1} = 1 + \frac{2}{qN-1}.$$

□

Remark 3.1.8 (Historical development)**.**
The upper bound $\sigma_{\max}(\boldsymbol{A})$ is already due to Selberg [100] since he proved a large sieve which is basically an inequality as in (3.1.12)*, where the constant in the upper bound only depends on the wrap-around distance and N, see also [7, p. 8]. Moitra used the Selberg minorant to prove the lower bound as presented in* (3.1.13) *but with a minorant for the characteristic function over the interval $[0, N-1]$, which leads to the sub-optimal lower bound $\sigma_{\min}(\boldsymbol{A}) \geq N - 1 - \frac{1}{q}$. This bound also does not cover the square Vandermonde matrix case with equispaced nodes. Aubel and Bölcskei improved this bound in [7] first by applying a generalized version of Hilbert's inequality from [81] to obtain $\sigma_{\min}(\boldsymbol{A}) \geq N - \frac{1}{q}$ (this is also stated in [89, Thm. 10.23]). Secondly, they use a dilation trick by Cohen to improve bounds for the more general case of nodes in the unit disc, resulting in Theorem 3.1.3. Taking that result and going back to the special case of nodes on the unit circle by letting the radius r tend to one in* (3.1.4) *yields the bound $\sigma_{\min}(\boldsymbol{A}) \geq N + \frac{1}{2} - \frac{1}{q}$, see [5, p. 22]. This bound is already valid for square Vandermonde matrices with equispaced nodes. Finally, Diederichs discovered in [34] that Moitra used a sub-optimal minorant and provided the bound in* (3.1.13)*.*

Remark 3.1.9 (Optimality)**.**
The inequality in (3.1.13) *produces a non-trivial lower bound as long as $q > \frac{1}{N+1}$. This is optimal in the following sense, cf. [34]. The calculation in* (3.1.13) *is independent of the number of nodes the Vandermonde matrix $\boldsymbol{A}$ consists of, as long as they are q-separated. In particular, for $q > \frac{1}{N+1}$ at most $M = N$ nodes are possible to place on the unit circle. If $q \leq \frac{1}{N+1}$ is allowed, $N+1$ q-separated nodes are possible to place on the unit circle. Thus, the calculation must apply to the Vandermonde matrix $\boldsymbol{A} \in \mathbb{C}^{M \times N}, M \geq N$, with these nodes, but clearly there exists a vector $0 \neq \boldsymbol{v} \in \mathbb{C}^M$ for which $\boldsymbol{A}^*\boldsymbol{v} = 0$. That means only the trivial bound is possible. In fact, only lower bounds of the form $\|\boldsymbol{A}\boldsymbol{u}\| \geq c\|\boldsymbol{u}\|^2$ for a constant c and all $\boldsymbol{u} \in \mathbb{C}^N$ would be effective for bounding the smallest singular value of $\boldsymbol{A}$ in this situation, see Theorem 2.1.15.*

We provide the following lower bound for the condition number. It shows that the upper bound for well-separated nodes is quite sharp.

Theorem 3.1.10 (Lower bound on the condition number)**.**
Let $\boldsymbol{A}$ be a Vandermonde matrix as in (3.1.6) *with with $N \geq M$ and q-separated nodes. Then we have*

$$\sigma_{\min}(\boldsymbol{A})^2 \leq N\left(1 - |d_{N-1}(q)|\right) \leq N \leq N\left(1 + |d_{N-1}(q)|\right) \leq \sigma_{\max}(\boldsymbol{A})^2.$$

In particular, we have

$$\operatorname{cond}(\boldsymbol{A})^2 \geq 1 + \frac{2}{\pi q N - 1}$$

for $qN \in \mathbb{N} + \frac{1}{2}$ almost matching the above upper bound from Theorem 3.1.7.

Proof. Without loss of generality, let $t_2 - t_1 = q$ since reordering rows does not change the condition number. Therefore, the entries of $\boldsymbol{A}\boldsymbol{A}^*$ are given by $N d_{N-1}(t_j - t_\ell)$ for $j, \ell = 1, \ldots, M$. We consider the upper left 2×2-block in

$$\boldsymbol{A}\boldsymbol{A}^* = \begin{pmatrix} \boldsymbol{C} & * \\ * & * \end{pmatrix}, \quad \boldsymbol{C} := N \begin{pmatrix} d_{N-1}(0) & d_{N-1}(q) \\ d_{N-1}(q) & d_{N-1}(0) \end{pmatrix}.$$

The eigenvalues of $\boldsymbol{C}$ are given by $\lambda_{\max}(\boldsymbol{C}) = N(d_{N-1}(0) + d_{N-1}(q))$ and $\lambda_{\min}(C) = N(d_{N-1}(0) - d_{N-1}(q))$. We apply Lemma 2.1.6 and get

$$\begin{aligned} \sigma_{\max}(\boldsymbol{A})^2 = \lambda_{\max}(\boldsymbol{A}\boldsymbol{A}^*) \geq \lambda_{\max}(\boldsymbol{C}) = N(d_{N-1}(0) + d_{N-1}(q)) \\ \sigma_{\min}(\boldsymbol{A})^2 = \lambda_{\min}(\boldsymbol{A}\boldsymbol{A}^*) \leq \lambda_{\min}(\boldsymbol{C}) = N(d_{N-1}(0) - d_{N-1}(q)). \end{aligned}$$

Lemma 2.4.11 provides $d_{N-1}(0) = 1$ and thus, we obtain

$$\operatorname{cond}(\boldsymbol{A})^2 = \frac{\lambda_{\max}(\boldsymbol{A}^*\boldsymbol{A})}{\lambda_{\min}(\boldsymbol{A}^*\boldsymbol{A})} \geq \frac{\lambda_{\max}(\boldsymbol{C})}{\lambda_{\min}(\boldsymbol{C})} = \frac{1 + |d_{N-1}(q)|}{1 - |d_{N-1}(q)|} = 1 + \frac{2\,|d_{N-1}(q)|}{1 - |d_{N-1}(q)|}. \tag{3.1.14}$$

By Definition 2.4.10 we have with $q = \frac{k+\frac{1}{2}}{N}$ and the inequality $\sin(x) \leq x$, that hold for all $x > 0$,

$$|d_{N-1}(q)| = \frac{\left|\sin\left((k+\frac{1}{2})\pi\right)\right|}{N\,|\sin(q\pi)|} = \frac{1}{N\,|\sin(q\pi)|} \geq \frac{1}{\pi q N}.$$

Hence, we can continue the above calculation and obtain

$$\operatorname{cond}(\boldsymbol{A})^2 = 1 + \frac{2\,|d_{N-1}(q)|}{1 - |d_{N-1}(q)|} = 1 + \frac{2}{1/\,|d_{N-1}(q)| - 1} \geq 1 + \frac{2}{\pi q N - 1}.$$

□

One question that may have been arisen is weather there are node sets, that lead to perfectly conditioned rectangular Vandermonde matrices. Indeed there are such sets and these are described by the following theorem.

Theorem 3.1.11 (Perfectly conditioned Vandermonde matrix)**.**
Let $\boldsymbol{A}$ be a Vandermonde matrix as in (3.1.6)*. Then we have*

$$\operatorname{cond}(\boldsymbol{A}) = 1$$

if and only if all nodes lie on a fixed grid with width $1/N$.

Proof. The proof uses the same idea stated in Section 3.1.1 from [18] for the case of quadratic Vandermonde matrices. Since $d_{N-1}(0) = 1$, Lemma 2.1.27 provides $\operatorname{cond}(\boldsymbol{A}) = 1$ if and only if $\boldsymbol{A}\boldsymbol{A}^* = N\boldsymbol{I}_M$. In particular the off diagonal entries are zero, i.e. necessarily we have $Nd_{N-1}(t-t') = 0$ for all distinct nodes $t \neq t'$. Where d_{N-1} is the modified Dirichlet kernel from Definition 2.4.10. Using its explicit form, we see that $d_{N-1}(t) = 0$ if and only if $t - t' \in \mathbb{Z}/N$. □

We end this section with a remark.

Remark 3.1.12 (Subsets of nodes)**.**
From the proof of Theorem 3.1.10 and in particular the application of the inclusion principle for eigenvalues of Hermitian matrices (Lemma 2.1.6), we can also derive the following simple principle. If we have a Vandermonde matrix corresponding to a set of nodes, then each Vandermonde matrix of the same degree, associated to a subset of nodes, has smaller or equal condition number. More precisely its smallest singular value at most rises and its largest singular value becomes at most smaller.

3.2 Clustered node configurations on the unit circle

In the previous section we have seen, if nodes are well-separated on the unit circle, then the accompanying Vandermonde matrices are well-conditioned. The upper bound from Theorem 3.1.7 provides appropriate results for nodes with minimal separation distance $q > \frac{1}{N}$. Theorems 3.1.1 and 3.1.2 already showed that the class of contiguous submatrices of the discrete Fourier matrix have an exponentially growing condition number. In particular, when identifying in Theorem 3.1.2 the inverse of the grid parameter $\frac{1}{\gamma}$ as the minimal separation distance q then we are in the situation where $q < 1/N$ and we obtain a first corollary.

Corollary 3.2.1 (Lower bound, contiguous submatrices of the Fourier matrix)**.**
Let $\boldsymbol{A} \in \mathbb{C}^{M\times N}$ a Vandermonde matrix with nodes consecutively lying on an equispaced fixed grid with parameter $\gamma \in \mathbb{N}_+, \gamma \leq N$ such that the minimal separation distance fulfills $q = 1/\gamma \leq \frac{2}{\mathrm{e}\pi N}$. Then we have

$$\operatorname{cond}(\boldsymbol{A}) \geq \frac{1}{4\sqrt{M}} \left(\frac{4}{\mathrm{e}\pi q N}\right)^{M-1}. \tag{3.2.1}$$

Proof. We simply apply Theorem 3.1.2 with the substitution $1/\gamma = q$ and use the condition on q to obtain

$$\operatorname{cond}(\boldsymbol{A}) \geq \frac{1 - \left(\frac{\mathrm{e}\pi q N}{4}\right)}{2\sqrt{M}} \left(\frac{4}{\mathrm{e}\pi q N}\right)^{M-1} \geq \frac{1}{4\sqrt{M}} \left(\frac{4}{\mathrm{e}\pi q N}\right)^{M-1}.$$

□

Also in the context of super-resolution of discrete measures with band-limited Fourier data [35, 31] found that the min-max error rate for the recovery of sparse objects in a super-resolution problem scales like SRF^{M-1} if $SRF > 1$, where SRF is the super-resolution factor given by the inverse of separation distance times the bandwidth of the given continuous Fourier data. It compares to $\frac{1}{qN}$ in our setting.

Taking these results together, there does not seem to be much hope for a relatively good conditioning of Vandermonde matrices with nodes that are not well-separated, especially, if there are a lot of nodes leading to a large exponent in the above lower bound. But having a closer look, in contiguous submatrices of the discrete Fourier matrix the separation between consecutive nodes are smaller than $1/N$ and the numerical experiment given in Figure 3.2.1 confirms the suspicion that if most nodes are well-separated, the situation becomes better. It suggest a condition number growing like $(qN)^{1-\lambda}$ with λ being the largest number of nodes that have consecutive separation smaller than $1/N$. We say that these nodes build a cluster (with a slight additional restriction). That means a priory knowledge about the geometric structure of the nodes set may help to establish bounds on the condition number more precisely. Therefore, we give a definition of clustered node configurations that orient at the "localized clumps" model from [73].

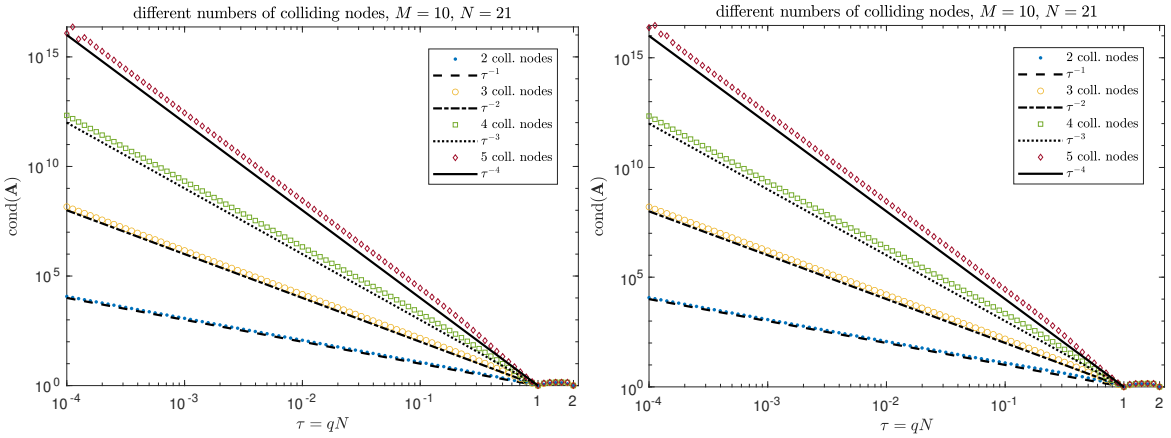


Figure 3.2.1: numerical example of $M = 10$ nodes initially given by $t_j = 2j/N, j = 0, \ldots, M-1$, with separation distance $q = 2/N$. Left: the first consecutive distances are decreased uniformly and the condition number of the corresponding Vandermonde matrix is calculated. We repeat this for different numbers of distances, i.e. different numbers of nodes colliding. For instance if we consider 3 colliding nodes, the first three nodes are scaled to qt_1, qt_2, qt_3, with new minimal separation $q = \tau/N < 2/N$. Right: same procedure as on the left, but every time we additionally decrease the distance between the 9th and 10th node such that $t_{10} = 18/N$ and $t_9 = t_{10} - q$, i.e. there is a second cluster with two nodes having minimal separation distance as well.

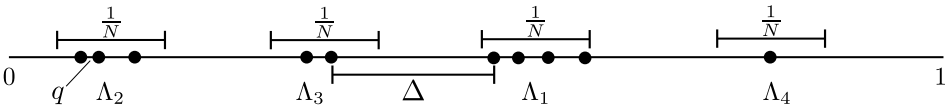


Figure 3.2.2: example of a clustered node configuration with $S = 4$ clusters and $\lambda = 4$.

Definition 3.2.2 (Clustered node configurations)**.**
A subset of nodes is called cluster *if it is contained in an interval of length* $1/N$*. For two clusters* $\Lambda', \Lambda'' \subset \Omega$*, we define*

$$\operatorname{dist}(\Lambda', \Lambda'') := \min\{\left|\boldsymbol{t}' - \boldsymbol{t}''\right|_{\mathbb{T}} : \ \boldsymbol{t}' \in \Lambda', \ \boldsymbol{t}'' \in \Lambda''\}.$$

The node set Ω *is called a* clustered node configuration *with* S *clusters if it can be written as*

$$\Omega = \bigcup_{l=1}^{S} \Lambda_l,$$

where the Λ_l *are clusters and the* minimal cluster separation Δ *fulfills*

$$\Delta := \min_{1 \le l < l' \le S} \operatorname{dist}(\Lambda_l, \Lambda_{l'}) > 1.$$

We order $|\Lambda_1| \ge |\Lambda_2| \ge \cdots \ge |\Lambda_S|$ *and denote the cardinality of the biggest cluster by* $\lambda := |\Lambda_1|$. *Figure 3.2.2 shows an illustration of an example. In passing, we note that the node set* Ω *is called* well-separated *with separation* Δ *if* $\lambda = 1$.

Remark 3.2.3 (Cluster)**.**
In order to simplify technical results later on and keep the number of parameters as low as possible, we define a cluster as a set of nodes contained in an interval of length $1/N$ *in contrast to [11] where an additional parameter for the cluster extent is used. This of course slightly reduces the class of node sets we cover with our results, as for instance Figure 3.2.1 suggests that nodes having consecutive distances below* $1/N$ *already build a cluster.*

Assuming the node set Ω to be a clustered node configuration, in [11] the first time an upper bound of the condition number of the corresponding Vandermonde matrix with exponent $\lambda - 1$, i.e.

$$\operatorname{cond}(\boldsymbol{A}) \le c(M) \left(\frac{1}{qN}\right)^{\lambda - 1}$$

with a constant depending only on M. This means that only the largest cluster determines the exponential growing which is good news when dealing with large node sets consisting of small clusters. Further results in that direction came up shortly after in [73, 12, 34]. We present these results in detail in Section 3.4.5 to have them at hand when comparing to our results from the next few sections. First, we start with the analysis of a special case, namely clustered node configurations with at most two nodes per cluster.

3.3 A Schur-complement technique for pair clusters

Let us consider a simple situation, where nodes on the unit circle are not well-separated anymore. We focus on two settings. In the first we take a well-separated node set and add a single node such that there is one cluster with a pair of nodes. The second is the more general setting with nodes consisting completely of clusters with two nodes. Most of this work is published in [68]. In this section we assume $N = 2n + 1, n \in \mathbb{N}$, unless stated differently. Results for Vandermonde matrices with even N can be obtained by using the monotony of singular values from Lemma 2.1.18. We consider the shifted Vandermonde matrix with degrees centered around the origin by

$$\operatorname{diag}(z_1^{-n}, \ldots, z_M^{-n})\boldsymbol{A} = \begin{pmatrix} z_1^{-n} & \cdots z_1^{-1} & 1 & z_1^{1} & \cdots & z_1^{n} \\ \vdots & & & \vdots & & \vdots \\ z_M^{-n} & \cdots z_M^{-1} & 1 & z_M^{1} & \cdots & z_M^{n} \end{pmatrix} \in \mathbb{C}^{M \times N}.$$

Since the singular values are unitary invariant (see Lemma 2.1.12), we have

$$\operatorname{cond}(\operatorname{diag}(z_1^{-n},\dots,z_M^{-n})\boldsymbol{A}) = \operatorname{cond}(\boldsymbol{A}).$$

Therefore, we can focus on studying the extremal eigenvalues of the Hermitian kernel matrix

$$\boldsymbol{K} := \operatorname{diag}(z_1^{-n},\dots,z_M^{-n})\boldsymbol{A}\boldsymbol{A}^* \operatorname{diag}(z_1^{-n},\dots,z_M^{-n})^* = \big(D_n\left(t_j - t_\ell\right)\big)_{j,\ell=1}^{M} \in \mathbb{C}^{M\times M} \tag{3.3.1}$$

that has as entries the Dirichlet kernel from Definition 2.4.5 evaluated at the node distances. The advantage of the Dirichlet kernel is, compared to the scaled modified Dirichlet kernel appearing in $\boldsymbol{A}\boldsymbol{A}^*$, that it is real and thus, dealing with derivatives is much easier. Using the monotony of the condition number with respect to node subsets, see Remark 3.1.12, a lower bound is effortlessly given. The lower bound serves also as a benchmark for the upper bound which to establish is more involved.

Lemma 3.3.1 (Lower Bound, pair clusters)**.**
Let $M, N = 2n+1 \in \mathbb{N}_+, M \geq N$ and $\boldsymbol{A} \in \mathbb{C}^{M\times N}$ as in (3.0.1) *a Vandermonde matrix associated to clustered node configuration containing a cluster of at least two nodes. Then we have*

$$\operatorname{cond}(\boldsymbol{A})^2 \geq \frac{12}{\pi^2(qN)^2} - 1 \geq \frac{1}{(qN)^2}$$

for $qN \leq \sqrt{12/\pi^2 - 1} \approx 0.46$ and $\operatorname{cond}(\boldsymbol{A})^2 \geq \frac{6}{\pi^2(qN)^2}$ for all $qN \leq 1$.

Proof. We proceed analogously to the proof of Theorem 3.1.10 and use $N\,|d_{N-1}(t)| = |D_n\left(t\right)|$, for all $t \in \mathbb{T}$, in (3.1.14). Together with $qN \leq 1$ we get

$$\operatorname{cond}(\boldsymbol{A})^2 \geq \frac{N + |D_n\left(q\right)|}{N - |D_n\left(q\right)|} = \frac{N + D_n\left(q\right)}{N - D_n\left(q\right)}.$$

Applying the bound $D_n\left(q\right) \geq N(1 - \frac{\pi^2}{6}(qN)^2)$ from Lemma 2.4.7 yields

$$\frac{N + D_n\left(q\right)}{N - D_n\left(q\right)} \geq \frac{2 - \frac{\pi^2}{6}(qN)^2}{\frac{\pi^2}{6}(qN)^2} = \frac{12}{\pi^2(qN)^2} - 1, \tag{3.3.2}$$

the first inequality of the assertion. If the second inequality should hold, we obtain by rearranging the condition

$$(qN)^2 \leq \frac{12}{\pi^2} - 1$$

which is satisfied as long as $qN \leq \sqrt{\frac{12}{\pi^2} - 1}$.

Finally, in (3.3.2) we can alternatively bound

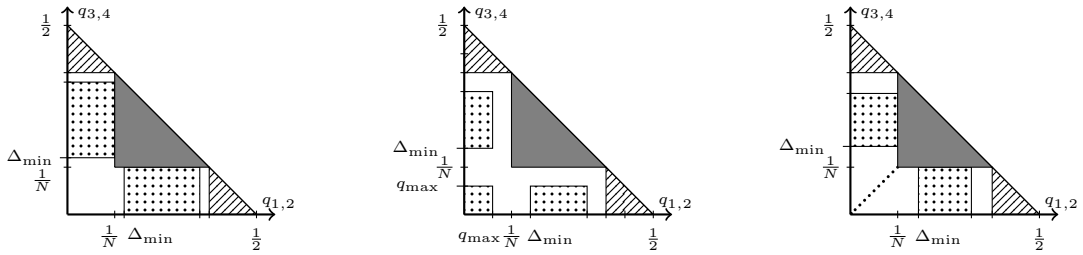
$$\frac{N + D_n\left(q\right)}{N - D_n\left(q\right)} \geq \frac{N}{N - D_n\left(q\right)} \geq \frac{6}{\pi^2(qN)^2}$$

for all $qN \leq 1$ and get the second bound for the condition number. □

The idea behind the technique presented in the next two sections, to obtain a lower bound on the smallest singular value and a larger bound on the largest singular value, is quiet simple. Since there are at most two nodes in each cluster it offers to partition the node set into

two subsets of well-separated nodes. Afterwards, we hope to be able to apply bounds for the extremal singular values of Vandermonde matrices with well separated-nodes we already know from Theorem 3.1.7. For the Hermitian kernel matrix $\boldsymbol{K}$ this means it gets partitioned into blocks and therefore, we can rephrase its inverse using a Schur-complement decomposition given in Lemma 2.1.21. Appearing norms are also bounded by Lemmata 2.1.20 and 2.1.22.

Figure 3.3.1 illustrates the situation for 4 nodes on the unit circle and when theorems from the next to sections are applicable. The parameter $q_{\max}$ and $\Delta_{\min}$ describe minimum separation distance of nodes and clusters respectively that are assumed in the theorems.



(a) Theorem 3.3.6 (b) Theorem 3.3.13 (c) Theorem 3.3.15

Figure 3.3.1: sketch of four-node configurations, $t_1 < t_2 < t_3 < t_4 \in [0,1)$, $t_1 = 0$, $t_3 = 1/2$, N large enough, $q_{1,2} := |t_1 - t_2|_{\mathbb{T}}$, $q_{3,4} := |t_3 - t_4|_{\mathbb{T}}$. Dotted: theorem can be applied; filled: well-separated; lined: 3 nodes building a cluster; empty areas: at most 2 nodes in a cluster, but not covered by results.

3.3.1 Nodes with one pair cluster

We consider a clustered node configuration with only one cluster of two nodes. This is the simplest situation coming from well-separated nodes sets and entering the regime of clustered node configurations. Let $M \geq 2$ and $0 = t_1 < \cdots < t_M \in \mathbb{T}$ such that

$$\begin{aligned} &|t_1 - t_2|_{\mathbb{T}} = q, && 0 < q \leq 1/N, \\ &|t_j - t_\ell|_{\mathbb{T}} \geq \Delta,\ j \neq \ell,\ \ell \geq 3, && 1/N < \Delta < \infty, \end{aligned} \tag{3.3.3}$$

see Figure 3.3.2 for an illustration. Due to periodicity, the choice $t_1 = 0$ and $|t_1 - t_2|_{\mathbb{T}} = q$ is without loss of generality.

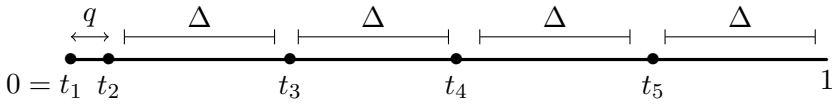


Figure 3.3.2: example of a node set with $M = 5$ satisfying (3.3.3).

Now, we estimate an upper bound on the condition number of the Hermitian matrix $\boldsymbol{K}$ by bounding $\|\boldsymbol{K}\|$ directly and applying Lemma 2.1.21 to $\boldsymbol{K}^{-1}$ before bounding $\left\|\boldsymbol{K}^{-1}\right\|$. For that, we introduce some notation.

Definition 3.3.2.
We define

$$\boldsymbol{a}_1 := \left(z_1^k\right)_{k\in\mathbb{Z},\,|k|\le n} \in \mathbb{C}^{1\times N} \quad \text{and} \quad \boldsymbol{A}_2 := \left(z_j^k\right)_{\substack{j=2,\dots,M\\ k\in\mathbb{Z},\,|k|\le n}} \in \mathbb{C}^{(M-1)\times N}$$

so that with

$$\boldsymbol{a}_1\boldsymbol{a}_1^* = N, \quad \boldsymbol{K}_2 = \boldsymbol{A}_2\boldsymbol{A}_2^* \in \mathbb{C}^{(M-1)\times(M-1)} \quad \text{and} \quad \boldsymbol{b} := \boldsymbol{A}_2\boldsymbol{a}_1^* = \begin{pmatrix} D_n(q) \\ D_n(t_3) \\ \vdots \\ D_n(t_M) \end{pmatrix} \in \mathbb{C}^{M-1}, \tag{3.3.4}$$

we have the partitioning

$$\operatorname{diag}(z_1^{-n},\dots,z_M^{-n})\boldsymbol{A} = \begin{pmatrix} \boldsymbol{a}_1 \\ \boldsymbol{A}_2 \end{pmatrix} \quad \text{and} \quad \boldsymbol{K} = \begin{pmatrix} N & \boldsymbol{b}^* \\ \boldsymbol{b} & \boldsymbol{K}_2 \end{pmatrix}, \tag{3.3.5}$$

where $\boldsymbol{A}_2$ is a shifted Vandermonde matrix with nodes that are at least Δ-separated.

Lemma 3.3.3.
Under the conditions of (3.3.3) *and for $\Delta N \ge 6$, we have*

$$\|\boldsymbol{K}\| \le 2.3N.$$

Proof. The key idea is to see the set of nodes as a union of two well-separated subsets and use the existing bounds for these. In contrast to the next section, here, one of the sets only consist of a single node. We start by noting that Lemma 2.1.14 and (3.3.4) yield $\|\boldsymbol{b}\|^2 \le \|\boldsymbol{a}_1\|^2\|\boldsymbol{A}_2\|^2 = N\|\boldsymbol{K}_2\|$. Together with the decomposition (3.3.5), the triangle inequality, Lemma 2.1.22, and Theorem 3.1.7, we obtain

$$\begin{aligned}\|\boldsymbol{K}\| &\le \left\| \begin{pmatrix} N & \boldsymbol{0}_{1\times(M-1)} \\ \boldsymbol{0}_{(M-1)\times 1} & \boldsymbol{K}_2 \end{pmatrix} \right\| + \left\| \begin{pmatrix} 0 & \boldsymbol{b}^* \\ \boldsymbol{b} & \boldsymbol{0}_{M-1} \end{pmatrix} \right\| \\ &\le \|\boldsymbol{K}_2\| + \|\boldsymbol{b}\| \le N\left(1 + \frac{1}{\Delta N} + \sqrt{1+\frac{1}{\Delta N}}\right).\end{aligned}$$

Finally, using the assumption $\Delta N \ge 6$ yields the numerical value. □

Lemma 3.3.4.
Under the conditions of (3.3.3) *and with $\boldsymbol{b}$ as in* (3.3.4)*, we have*

$$\boldsymbol{b} = \boldsymbol{K}_2\boldsymbol{e}_1 + \boldsymbol{r},$$

where $\boldsymbol{e}_1 \in \mathbb{R}^{(M-1)}$ denotes the first unit vector and $\boldsymbol{r} = (r_1,\dots,r_{M-1})^\top \in \mathbb{R}^{M-1}$ satisfies

$$\|\boldsymbol{r}\|^2 \le (N - D_n(q))^2 + N^2(qN)^2\left(\frac{\pi^4}{12(\Delta N)^2} + \frac{1.21\pi}{(\Delta N)^3} + \frac{\pi^4}{180(\Delta N)^4}\right).$$

Proof. The vector $\boldsymbol{b}$ can be approximated by the first column of $\boldsymbol{K}_2$ in the sense that

$$\boldsymbol{b} = \begin{pmatrix} D_n(q) \\ D_n(t_3) \\ \vdots \\ D_n(t_M) \end{pmatrix} = \begin{pmatrix} D_n(0) \\ D_n(t_3 - q) \\ \vdots \\ D_n(t_M - q) \end{pmatrix} + \begin{pmatrix} r_1 \\ \vdots \\ r_{M-1} \end{pmatrix}.$$

We have $|r_1| = N - D_n(q)$ and for $j = 2, \ldots, M-1$ the mean value theorem yields

$$|r_j| = |D_n(t_{j+1}) - D_n(t_{j+1} - q)| = \left|D_n'(\xi_j)\right| q, \quad \xi_j \in (|t_{j+1} - q|_{\mathbb{T}}, |t_{j+1}|_{\mathbb{T}}).$$

Note that, in the worst case, half of the nodes can be as close as possible (under the assumed separation condition) to t_2 not only on its right but also on its left. Hence, for $j = 2, \ldots, \lceil \frac{M}{2} \rceil$, $\xi_j \geq (j-1)\Delta$ and Lemma 2.4.7 lead to

$$|r_j| \leq \left(\frac{\pi}{2N|\xi_j|} + \frac{1}{2N^2|\xi_j|^2} \right) qN^2 \leq \left(\frac{\pi}{2(j-1)(\Delta N)} + \frac{1}{2(j-1)^2(\Delta N)^2} \right) qN^2.$$

Thus, for all nodes, we get

$$\sum_{j=2}^{M-1} |r_j|^2 \leq 2 \sum_{j=2}^{\lceil M/2 \rceil} |r_j|^2 \leq q^2 N^4 \left(\frac{\pi^2}{2(\Delta N)^2} \underbrace{\sum_{j=1}^{\infty} \frac{1}{j^2}}_{=\frac{\pi^2}{6}} + \frac{\pi}{(\Delta N)^3} \underbrace{\sum_{j=1}^{\infty} \frac{1}{j^3}}_{\leq 1.21} + \frac{1}{2(\Delta N)^4} \underbrace{\sum_{j=1}^{\infty} \frac{1}{j^4}}_{=\frac{\pi^4}{90}} \right). \tag{3.3.6}$$

□

Lemma 3.3.5.
Under the conditions of (3.3.3) *and for* $\Delta N \geq 5$*, we have*

$$\left\|\boldsymbol{K}^{-1}\right\| \leq \frac{C(\Delta N)}{N(qN)^2},$$

where

$$C(\Delta N) = \left(\frac{2\Delta N - 1}{\Delta N - 1} + \sqrt{\frac{\Delta N}{\Delta N - 1}} \right) \cdot \left[2 - \frac{\Delta N}{\Delta N - 1} \left(1 + \frac{\pi^4}{12(\Delta N)^2} + \frac{1.21\pi}{(\Delta N)^3} + \frac{\pi^4}{180(\Delta N)^4} \right) \right]^{-1}.$$

is positive.

Proof. We consider $\boldsymbol{K}$ decomposed as in (3.3.5) and apply Lemma 2.1.21 with respect to $\boldsymbol{K}_2$ to obtain

$$\boldsymbol{K}^{-1} = \begin{pmatrix} 1 & \boldsymbol{0}_{1\times(M-1)} \\ -\boldsymbol{K}_2^{-1}\boldsymbol{b} & \boldsymbol{I}_{M-1} \end{pmatrix} \begin{pmatrix} (N - \boldsymbol{b}^*\boldsymbol{K}_2^{-1}\boldsymbol{b})^{-1} & \boldsymbol{0}_{1\times(M-1)} \\ \boldsymbol{0}_{(M-1)\times 1} & \boldsymbol{K}_2^{-1} \end{pmatrix} \begin{pmatrix} 1 & -\boldsymbol{b}^*\boldsymbol{K}_2^{-1} \\ \boldsymbol{0}_{(M-1)\times 1} & \boldsymbol{I}_{M-1} \end{pmatrix}$$

and thus,

$$\left\|\boldsymbol{K}^{-1}\right\| \leq \left\|\begin{pmatrix} 1 & \boldsymbol{0}_{1\times(M-1)} \\ -\boldsymbol{K}_2^{-1}\boldsymbol{b} & \boldsymbol{I}_{M-1}\end{pmatrix}\right\|^2 \max\left\{\left\|\boldsymbol{K}_2^{-1}\right\|, \left|\left(N-\boldsymbol{b}^*\boldsymbol{K}_2^{-1}\boldsymbol{b}\right)^{-1}\right|\right\}.$$

First of all, we establish an upper bound for the norm of the triangular matrix. Since $\boldsymbol{A}_2$ has full row rank as shifted Vandermonde matrix of distinct nodes, Lemma 2.1.25 shows $\left((\boldsymbol{A}_2\boldsymbol{A}_2^*)^{-1}\boldsymbol{A}_2\right)^* = \boldsymbol{A}_2^\dagger$. Then (3.3.4) and Theorem 3.1.7 imply

$$\left\|\boldsymbol{K}_2^{-1}\boldsymbol{b}\right\| = \left\|(\boldsymbol{A}_2\boldsymbol{A}_2^*)^{-1}\boldsymbol{A}_2\boldsymbol{a}_1^*\right\| \leq \left\|\boldsymbol{A}_2^\dagger\right\| \|\boldsymbol{a}_1\| \leq \sqrt{\frac{\Delta N}{\Delta N-1}}.$$

Together with Lemma 2.1.22, we obtain

$$\left\|\begin{pmatrix} 1 & \boldsymbol{0}_{1\times(M-1)} \\ -\boldsymbol{K}_2^{-1}\boldsymbol{b} & \boldsymbol{I}_{M-1}\end{pmatrix}\right\|^2 \leq 1 + \left\|\boldsymbol{K}_2^{-1}\boldsymbol{b}\right\| + \left\|\boldsymbol{K}_2^{-1}\boldsymbol{b}\right\|^2 \leq \frac{2\Delta N-1}{\Delta N-1} + \sqrt{\frac{\Delta N}{\Delta N-1}}. \quad (3.3.7)$$

The next step is to bound $(N-\boldsymbol{b}^*\boldsymbol{K}_2^{-1}\boldsymbol{b})^{-1}$. Lemma 3.3.4 yields

$$\boldsymbol{b}^*\boldsymbol{K}_2^{-1}\boldsymbol{b} = (\boldsymbol{K}_2\boldsymbol{e}_1+\boldsymbol{r})^*\boldsymbol{K}_2^{-1}(\boldsymbol{K}_2\boldsymbol{e}_1+\boldsymbol{r}) = 2D_n(q) - D_n(0) + \boldsymbol{r}^*\boldsymbol{K}_2^{-1}\boldsymbol{r}.$$

Applying the second part of Lemma 3.3.4, Lemma 2.4.7, and Theorem 3.1.7 yields

$$\begin{aligned}
N-\boldsymbol{b}^*\boldsymbol{K}_2^{-1}\boldsymbol{b} &\geq 2\left(N-D_n(q)\right) - \|\boldsymbol{r}\|^2\left\|\boldsymbol{K}_2^{-1}\right\| \\
&\geq \left(N-D_n(q)\right)\left(2-\left(N-D_n(q)\right)\left\|\boldsymbol{K}_2^{-1}\right\|\right) - \left\|\boldsymbol{K}_2^{-1}\right\|\sum_{j=2}^{M-1}|r_j|^2 \\
&\geq N(qN)^2\left(2-N\left\|\boldsymbol{K}_2^{-1}\right\|\right) - \|\boldsymbol{K}_2^{-1}\|N^2(qN)^2\left(\frac{\pi^4}{12(\Delta N)^2}+\frac{1.21\pi}{(\Delta N)^3}+\frac{\pi^4}{180(\Delta N)^4}\right) \\
&\geq N(qN)^2\left[2-\frac{\Delta N}{\Delta N-1}\left(1+\frac{\pi^4}{12(\Delta N)^2}+\frac{1.21\pi}{(\Delta N)^3}+\frac{\pi^4}{180(\Delta N)^4}\right)\right].
\end{aligned}$$

For $\Delta N \geq 5$, the most inner bracketed term takes values in $(1, 1.4)$ such that the square bracketed term is positive. Forming the reciprocal gives the result, since Theorem 3.1.7 also implies

$$N\left\|\boldsymbol{K}_2^{-1}\right\| \leq \frac{\Delta N}{\Delta N-1} \leq \frac{\Delta N-1}{\Delta N-2} \leq \left[2-\frac{\Delta N}{\Delta N-1}\left(1+\ldots\right)\right]^{-1}. \quad (3.3.8)$$

□

Theorem 3.3.6 (Upper bound)**.**
Under the conditions of (3.3.3) *with* $\Delta N \geq \Delta_{\min} N = 6$*, we have*

$$\operatorname{cond}(\boldsymbol{A}) \leq \frac{4}{qN}.$$

Proof. The bound follows from Lemmata 3.3.3 and 3.3.5 with $C(\Delta N) \leq C(6) \leq 6.5$. □

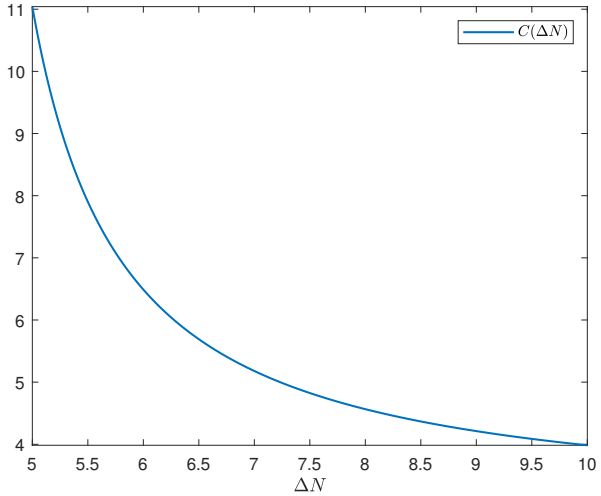


Figure 3.3.3: $C(\Delta N)$ from Lemma 3.3.5.

Lower and upper bounds in Lemma 3.3.1 and Theorem 3.3.6 yield

$$\frac{1}{qN} \leq \operatorname{cond}(\boldsymbol{A}) \leq \frac{4}{qN}$$

for $qN \leq 0.46$ and $6 \leq \Delta N$. The condition on Δ implies that for specific configurations of M nodes, our result becomes effective as early as $N \approx 6M$.

Remark 3.3.7 (Constants)**.**
Some comments regarding what is lost during our proof:

i) The constant in Lemma 3.3.3 is a numerical value for all $\Delta N \geq 6$, indeed the proof is valid for all values $\Delta N > 1$. The case $M = 2$ shows that Lemmata 3.3.3 and 3.3.4 are reasonable sharp since in that case $\|\boldsymbol{K}\| = N + D_n(q) \geq N(2 - \pi^2(qN)^2/6)$ and $\|\boldsymbol{r}\| = N - D_n(q) \geq N(qN)^2$, see Lemma 2.4.7 for the two inequalities.

ii) In Lemma 3.3.5, the constant $C(\Delta N)$ is monotone decreasing in ΔN, see also Figure 3.3.3. It is bounded below by 3 which is due to the relatively crude norm estimate on the block triangular factors in the Schur-complement decomposition. Note that the left hand side in (3.3.7) is bounded from below by $1 + \left\|\boldsymbol{K}_2^{-1}\boldsymbol{b}\right\|^2$. This can be seen by the following. Let $\boldsymbol{x} \in \mathbb{C}^{M-1}$ with $\|\boldsymbol{x}\| = 1$ such that we have $(\boldsymbol{K}_2^{-1}\boldsymbol{b})^*\boldsymbol{x} = \left\|\boldsymbol{K}_2^{-1}\boldsymbol{b}\right\|$. Then we have, after extending $\boldsymbol{x}$ by adding a zero, again by the variation characterization of singular values

$$\begin{aligned}\left\|\begin{pmatrix} 1 & \mathbf{0}_{1\times(M-1)} \\ -\boldsymbol{K}_2^{-1}\boldsymbol{b} & \boldsymbol{I}_{M-1}\end{pmatrix}\right\|^2 &\geq \left\|\begin{pmatrix} 1 & (-\boldsymbol{K}_2^{-1}\boldsymbol{b})^* \\ \mathbf{0}_{(M-1)\times 1} & \boldsymbol{I}_{M-1}\end{pmatrix}\begin{pmatrix} 0 \\ \boldsymbol{x}\end{pmatrix}\right\|^2 \\ &= \left((\boldsymbol{K}_2^{-1}\boldsymbol{b})^*\boldsymbol{x}\right)^2 + \|\boldsymbol{x}\|^2 = 1 + \left\|\boldsymbol{K}_2^{-1}\boldsymbol{b}\right\|^2.\end{aligned}$$

An additional minor improvement on $C(\Delta N)$ and on the range of admissible values for Δ can be achieved when treating the lower bound of $N - \boldsymbol{b}^*\boldsymbol{K}_2^{-1}\boldsymbol{b}$ in the proof of Lemma 3.3.5 as quadratic function in $N - D_n(q)$ and applying lower bounds for $N - D_n(q)$ simultaneously.

3.3.2 Several pair clusters

We now study the situation in which the Vandermonde matrix has nodes consisting of pair clusters. Let $n \in \mathbb{N}$, $N = 2n+1$, $c \geq 1$ and let $t_1 < \cdots < t_{\frac{M}{2}} \in \mathbb{T}$ and $t_{\frac{M}{2}+1} < \cdots < t_M \in \mathbb{T}$ for $M \geq 4$ even such that

$$\begin{aligned} q \leq \left|t_j - t_{j+\frac{M}{2}}\right|_{\mathbb{T}} \leq cq, \quad j = 1, \ldots, \frac{M}{2}, \qquad 0 < cq \leq 1/N, \\ \Delta \leq |t_j - t_\ell|_{\mathbb{T}}, \quad j < \ell, \ell \neq j + \frac{M}{2}, \qquad 1/N < \Delta < \infty, \end{aligned} \tag{3.3.9}$$

then $\Omega = \{t_1, \ldots, t_M\}$ is a clustered node configuration with pair clusters, see Figure 3.3.4 for an illustration. The constant c measures the uniformity of the clusters. For subsequent use, we additionally introduce the following wrap-around distance of indices $|j-\ell|' := \min_{r\in\mathbb{Z}} \left|j - \ell + r\frac{M}{2}\right|$ with respect to $\frac{M}{2}$.

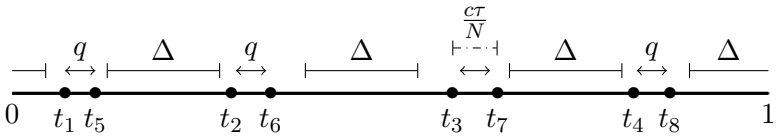


Figure 3.3.4: example of a node set with $M = 8$ satisfying (3.3.9).

Definition 3.3.8.
We define

$$\boldsymbol{A}_1 := \left(z_j^k\right)_{\substack{j=1,\ldots,M/2 \\ k\in\mathbb{Z},\,|k|\leq n}} \in \mathbb{C}^{(M/2)\times N} \quad \text{and} \quad \boldsymbol{A}_2 := \left(z_j^k\right)_{\substack{j=M/2+1,\ldots,M \\ k\in\mathbb{Z},\,|k|\leq n}} \in \mathbb{C}^{(M/2)\times N}$$

so that with $\boldsymbol{K}_1 := \boldsymbol{A}_1\boldsymbol{A}_1^*$, $\boldsymbol{K}_2 := \boldsymbol{A}_2\boldsymbol{A}_2^*$ *and* $\boldsymbol{B} := \boldsymbol{A}_2\boldsymbol{A}_1^*$ *all in* $\mathbb{C}^{(M/2)\times(M/2)}$ *we have the partitioning*

$$\operatorname{diag}(z_1^{-n}, \ldots, z_M^{-n})\boldsymbol{A} = \begin{pmatrix} \boldsymbol{A}_1 \\ \boldsymbol{A}_2 \end{pmatrix}, \quad \boldsymbol{K} = \begin{pmatrix} \boldsymbol{K}_1 & \boldsymbol{B}^* \\ \boldsymbol{B} & \boldsymbol{K}_2 \end{pmatrix}. \tag{3.3.10}$$

Note that under the assumptions in (3.3.9) *the shifted Vandermonde matrices* $\boldsymbol{A}_1$ *and* $\boldsymbol{A}_2$ *are each corresponding to nodes that are at least* Δ*-separated.*

The proof technique we use is analogous to the one we used in the case of nodes with only one pair cluster. The difference is that we have a matrix $\boldsymbol{K}_1$ instead of a scalar and the block $\boldsymbol{B}$ is a matrix instead of a vector. Subsequently, Lemma 3.3.9 establishes an upper bound on $\|\boldsymbol{K}\|$ and Lemmata 3.3.10 to 3.3.12 establish an upper bound on $\left\|\boldsymbol{K}^{-1}\right\|$.

Lemma 3.3.9.
Under the conditions of (3.3.9), *we have*

$$\|\boldsymbol{K}\| \leq 2N \cdot \frac{\Delta N + 1}{\Delta N}.$$

Proof. Similar to Lemma 3.3.3, we start by noting that $\|\boldsymbol{B}\|^2 \leq \|\boldsymbol{K}_1\| \, \|\boldsymbol{K}_2\|$. Together with the decomposition (3.3.10), the triangle inequality, Lemma 2.1.22, and Theorem 3.1.7, this leads to

$$\begin{aligned}\|\boldsymbol{K}\| &\leq \left\| \begin{pmatrix} \boldsymbol{K}_1 & \boldsymbol{0}_{M/2} \\ \boldsymbol{0}_{M/2} & \boldsymbol{K}_2 \end{pmatrix} \right\| + \left\| \begin{pmatrix} \boldsymbol{0}_{M/2} & \boldsymbol{B}^* \\ \boldsymbol{B} & \boldsymbol{0}_{M/2} \end{pmatrix} \right\| \\ &\leq \max\left\{ \|\boldsymbol{K}_1\|, \|\boldsymbol{K}_2\| \right\} + \sqrt{\|\boldsymbol{K}_1\| \, \|\boldsymbol{K}_2\|} \leq 2N \cdot \frac{\Delta N + 1}{\Delta N}.\end{aligned}$$

□

Lemma 3.3.10.
Under the conditions of (3.3.9), $\boldsymbol{R}_1 := \boldsymbol{B} - \boldsymbol{K}_1$ *fulfills*

$$\|\boldsymbol{R}_1\| \leq N - D_n\left(cq\right) + cqN^2 \left(\frac{\pi(\log\left\lfloor \frac{M}{4} \right\rfloor + 1)}{\Delta N} + \frac{\pi^2}{6(\Delta N)^2} \right).$$

Proof. The Dirichlet kernel D_n is monotone decreasing on $[0, 1/N]$. Hence, for the diagonal entries we obtain

$$|(\boldsymbol{R}_1)_{jj}| = \left| D_n\left(t_j - t_{j+\frac{M}{2}}\right) - N \right| = N - D_n\left(t_j - t_{j+\frac{M}{2}}\right) \leq N - D_n\left(cq\right).$$

The off diagonal entries are bounded by the mean value theorem and Lemma 2.4.7 as

$$\begin{aligned}|(\boldsymbol{R}_1)_{j\ell}| &= \left| D_n\left(t_j - t_\ell\right) - D_n\left(t_{j+\frac{M}{2}} - t_\ell\right) \right| \\ &\leq \left| D_n'\left(\xi_{j\ell}\right) \right| cq \leq c(qN)N^2 \left(\frac{\pi}{2N\xi_{j\ell}} + \frac{1}{2N^2\xi_{j\ell}^2} \right),\end{aligned}$$

where $\left(\left| t_{j+\frac{M}{2}} - t_\ell \right|_{\mathbb{T}}, |t_j - t_\ell|_{\mathbb{T}} \right) \ni \xi_{j\ell} \geq |j-\ell|' \, \Delta$ implies

$$|(\boldsymbol{R}_1)_{j\ell}| \leq cqN^2 \left(\frac{\pi}{2(\Delta N)\,|j-\ell|'} + \frac{1}{2(\Delta N)^2(|j-\ell|')^2} \right) =: (\widetilde{\boldsymbol{R}}_1)_{j\ell}$$

for $j, \ell = 1, \ldots, \frac{M}{2}$, $j \neq \ell$. Additionally, we set $(\widetilde{\boldsymbol{R}}_1)_{jj} := N - D_n\left(cq\right)$. We bound the spectral norm of $\boldsymbol{R}_1$ by the one of the real symmetric matrix $\widetilde{\boldsymbol{R}}_1$ using Lemma 2.1.23 and proceed by

$$\|\boldsymbol{R}_1\| \leq \left\| \widetilde{\boldsymbol{R}}_1 \right\| \leq \left\| \widetilde{\boldsymbol{R}}_1 \right\|_\infty \leq N - D_n\left(cq\right) + 2c(qN)N \sum_{j=1}^{\lfloor \frac{M}{4} \rfloor} \left(\frac{\pi}{2j(\Delta N)} + \frac{1}{2j^2(\Delta N)^2} \right),$$

from which the assertion follows since the first sum can be bounded by means of

$$1 + \int_1^{\lfloor \frac{M}{4} \rfloor} \frac{1}{x} \, dx = 1 + \log\left\lfloor \frac{M}{4} \right\rfloor.$$

□

Lemma 3.3.11.
Under the conditions of (3.3.9), $\boldsymbol{R}_1 = \boldsymbol{B} - \boldsymbol{K}_1$ *and* $\boldsymbol{R}_2 := \boldsymbol{B} - \boldsymbol{K}_2$ *fulfill*

$$\left\|2N\boldsymbol{I}_{M/2} + \boldsymbol{R}_1^* + \boldsymbol{R}_2\right\| \leq 2D_n\left(q\right) + Nc^2(qN)^2\left(\frac{\pi^2(\log\lfloor\frac{M}{4}\rfloor + 1)}{\Delta N} + \frac{\pi^3}{3(\Delta N)^2} + \frac{2.42}{(\Delta N)^3}\right).$$

Proof. First, note that

$$(\boldsymbol{R}_1^* + \boldsymbol{R}_2)_{j\ell} = D_n\left(t_{j+\frac{M}{2}} - t_\ell\right) + D_n\left(t_j - t_{\ell+\frac{M}{2}}\right) - D_n\left(t_{j+\frac{M}{2}} - t_{\ell+\frac{M}{2}}\right) - D_n\left(t_j - t_\ell\right).$$

Monotonicity of the Dirichlet kernel D_n on $t \in [0, 1/N]$ gives

$$\left|(2N\boldsymbol{I}_{M/2} + \boldsymbol{R}_1^* + \boldsymbol{R}_2)_{jj}\right| = 2\left|D_n\left(t_{j+\frac{M}{2}} - t_j\right)\right| \leq 2D_n\left(q\right)$$

for $j = \ell$. For each fixed off diagonal entry $j \neq \ell$, the matrix $2N\boldsymbol{I}_{M/2}$ has no contribution. We write the node $t_{j+M/2}$ as a perturbation of t_j by $h_j := t_{j+M/2} - t_j$ and expand the Dirichlet kernel by its Taylor polynomial of degree 2 in the point $\hat{h} := t_j - t_\ell + \frac{h_j - h_\ell}{2}$. Using

$$D_n\left(h\right) = D_n(\hat{h}) + D_n'(\hat{h})(h - \hat{h}) + \frac{D_n''(\xi)}{2}(h - \hat{h})^2$$

for some $\xi \in [\hat{h}, h] \cup [h, \hat{h}]$, the constant term as well as the linear term cancel out and we get

$$\begin{aligned}
&D_n(t_j + h_j - t_\ell) + D_n(t_j - t_\ell - h_\ell) - D_n(t_j + h_j - t_\ell - h_\ell) - D_n(t_j - t_\ell)\\
&= \frac{1}{8}\left(D_n''(\xi_1)(h_j + h_\ell)^2 + D_n''(\xi_2)(h_j + h_\ell)^2 + D_n''(\xi_3)(h_j - h_\ell)^2 + D_n''(\xi_4)(h_j - h_\ell)^2\right).
\end{aligned}$$

Lemma 2.4.7 and $\xi_1, \ldots, \xi_4 \geq |j - \ell|'\,\Delta$ imply

$$\begin{aligned}
|(\boldsymbol{R}_1^* + \boldsymbol{R}_2)_{j\ell}| \leq &\frac{N^3}{4}\left(\frac{\pi^2}{2\,|j - \ell|'\,(\Delta N)} + \frac{\pi}{(|j - \ell|')^2(\Delta N)^2} + \frac{1}{(|j - \ell|')^3(\Delta N)^3}\right)\\
&\cdot\left((h_j + h_\ell)^2 + (h_j - h_\ell)^2\right)
\end{aligned}$$

and hence by $h_j, h_\ell \leq cq$

$$\begin{aligned}
&\left|(2N\boldsymbol{I}_{M/2} + \boldsymbol{R}_1^* + \boldsymbol{R}_2)_{j\ell}\right|\\
&\leq c^2(qN)^2N\left(\frac{\pi^2}{2\,|j - \ell|'\,(\Delta N)} + \frac{\pi}{(|j - \ell|')^2(\Delta N)^2} + \frac{1}{(|j - \ell|')^3(\Delta N)^3}\right).
\end{aligned}$$

The matrix $2N\boldsymbol{I}_{M/2} + \boldsymbol{R}_1^* + \boldsymbol{R}_2$ is real symmetric so that

$$\begin{aligned}
\left\|2N\boldsymbol{I}_{M/2} + \boldsymbol{R}_1^* + \boldsymbol{R}_2\right\| &\leq \left\|2N\boldsymbol{I}_{M/2} + \boldsymbol{R}_1^* + \boldsymbol{R}_2\right\|_\infty\\
&\leq 2D_n\left(q\right) + 2\sum_{j=1}^{\lfloor\frac{M}{4}\rfloor} c^2(qN)^2N\left(\frac{\pi^2}{2j(\Delta N)} + \frac{\pi}{j^2(\Delta N)^2} + \frac{1}{j^3(\Delta N)^3}\right)\\
&\leq 2D_n\left(q\right) + 2c^2(qN)^2N\left(\frac{\pi^2(\log\lfloor\frac{M}{4}\rfloor + 1)}{2(\Delta N)} + \frac{\pi^3}{6(\Delta N)^2} + \frac{1.21}{(\Delta N)^3}\right)
\end{aligned}$$

and therefore the result holds. □

Lemma 3.3.12.
Under the conditions of (3.3.9) *with* $qN \leq \frac{1}{2}$ *and* $\Delta N \geq 2$, *such that*

$$\widetilde{C}(qN, \Delta N, c, M) := 2 - \frac{c^2\pi^2(\log\lfloor\frac{M}{4}\rfloor + 1)}{(\Delta N)} - \frac{c^2\pi^3}{3(\Delta N)^2} - \frac{2.42c^2}{(\Delta N)^3} - \frac{\Delta N}{\Delta N - 1}\left(\frac{c^2\pi^2}{6}(qN) + \frac{c\pi(\log\lfloor\frac{M}{4}\rfloor + 1)}{\Delta N} + \frac{c\pi^2}{6(\Delta N)^2}\right)^2$$

is positive, we have

$$\left\|\boldsymbol{K}^{-1}\right\| \leq \frac{C(qN, \Delta N, c, M)}{N(qN)^2},$$

where

$$C(qN, \Delta N, c, M) := \left(\frac{2\Delta N}{\Delta N - 1} + \sqrt{\frac{\Delta N + 1}{\Delta N - 1}}\right) / \widetilde{C}(qN, \Delta N, c, M).$$

Figure 3.3.5 visualizes the values of the constant $\widetilde{C}(qN, \Delta N, c, M)$ *with respect to* Δ *and* q. *Please note that i) increasing the constant* c *by a factor* $\sqrt{2}$ *has to be compensated approximately by halving* q *and doubling* Δ *and ii) increasing the number of nodes* M *from* 4 *to* 64 *has to be compensated approximately by tripling* Δ.

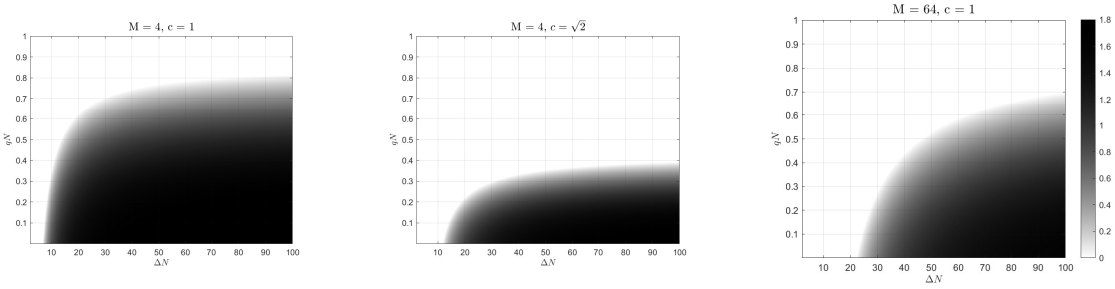


Figure 3.3.5: values of $\widetilde{C}(qN, \Delta N, c, M)$ in Lemma 3.3.12 depending on qN and ΔN for different M and c. Negative values are set to zero.

Proof. We proceed analogously to Lemma 3.3.5 and apply Lemma 2.1.21 to the matrix $\boldsymbol{K}$ decomposed as in (3.3.10) and obtain

$$\|\boldsymbol{K}^{-1}\| \leq \max\{\|\boldsymbol{K}_1^{-1}\|, \|(\boldsymbol{K}_2 - \boldsymbol{B}\boldsymbol{K}_1^{-1}\boldsymbol{B}^*)^{-1}\|\} \left\|\begin{pmatrix} \boldsymbol{I}_{M/2} & \boldsymbol{0}_{M/2} \\ -\boldsymbol{B}\boldsymbol{K}_1^{-1} & \boldsymbol{I}_{M/2}\end{pmatrix}\right\|^2. \tag{3.3.11}$$

Definition 3.3.8 and Theorem 3.1.7 yield

$$\left\|\boldsymbol{B}\boldsymbol{K}_1^{-1}\right\| \leq \|\boldsymbol{A}_2\| \left\|\boldsymbol{A}_1^{\dagger}\right\| \leq \sqrt{\frac{\Delta N + 1}{\Delta N - 1}},$$

together with Lemma 2.1.22, we obtain

$$\left\|\begin{pmatrix} \boldsymbol{I}_{M/2} & \boldsymbol{0}_{M/2} \\ -\boldsymbol{B}\boldsymbol{K}_1^{-1} & \boldsymbol{I}_{M/2}\end{pmatrix}\right\|^2 \leq 1 + \left\|\boldsymbol{B}\boldsymbol{K}_1^{-1}\right\| + \left\|\boldsymbol{B}\boldsymbol{K}_1^{-1}\right\|^2 \leq \frac{2\Delta N}{\Delta N - 1} + \sqrt{\frac{\Delta N + 1}{\Delta N - 1}}.$$

Now, we estimate $\left\|(\boldsymbol{K}_2 - \boldsymbol{B}\boldsymbol{K}_1^{-1}\boldsymbol{B}^*)^{-1}\right\|$, which is done by the following steps:

i) First, note that $\boldsymbol{I}_{M/2} - \boldsymbol{A}_1^\dagger \boldsymbol{A}_1$ is an orthogonal projector with norm one (directly seen by looking at the SVD of $\boldsymbol{A}$) and thus Theorem 3.1.7 implies

$$\left\|\boldsymbol{K}_2 - \boldsymbol{B}\boldsymbol{K}_1^{-1}\boldsymbol{B}^*\right\| \leq \|\boldsymbol{A}_2\| \left\|\boldsymbol{I}_{M/2} - \boldsymbol{A}_1^\dagger \boldsymbol{A}_1\right\| \|\boldsymbol{A}_2^*\| \leq \|\boldsymbol{A}_2\|^2 < 2N.$$

We apply Lemma 2.1.20 with $\eta = 2N$, use the identities $\boldsymbol{R}_1 = \boldsymbol{B} - \boldsymbol{K}_1$ and $\boldsymbol{R}_2 = \boldsymbol{B} - \boldsymbol{K}_2$, apply the triangular inequality, and the sub-multiplicativity of the matrix norm to get

$$\begin{aligned}\left\|(\boldsymbol{K}_2 - \boldsymbol{B}\boldsymbol{K}_1^{-1}\boldsymbol{B}^*)^{-1}\right\| &= \frac{1}{2N - \left\|2N\boldsymbol{I}_{M/2} - \boldsymbol{K}_2 + \boldsymbol{B}\boldsymbol{K}_1^{-1}\boldsymbol{B}^*\right\|} \\ &\leq \frac{1}{2N - \left\|2N\boldsymbol{I}_{M/2} + \boldsymbol{R}_1^* + \boldsymbol{R}_2\right\| - \|\boldsymbol{R}_1\|^2 \left\|\boldsymbol{K}_1^{-1}\right\|}.\end{aligned} \tag{3.3.12}$$

ii) Lemma 3.3.11 leads to

$$\begin{aligned}2N - \left\|2N\boldsymbol{I}_{M/2} + \boldsymbol{R}_1^* + \boldsymbol{R}_2\right\| \geq\ & 2(N - D_n(q)) \\ & - c^2(qN)^2 N \left(\frac{\pi^2(\log\lfloor\frac{M}{4}\rfloor + 1)}{\Delta N} + \frac{\pi^3}{3(\Delta N)^2} + \frac{2.42}{(\Delta N)^3}\right).\end{aligned}$$

iii) We apply Theorem 3.1.7 and Lemma 3.3.10 to get

$$\begin{aligned}&\|\boldsymbol{R}_1\|^2 \left\|\boldsymbol{K}_1^{-1}\right\| \\ &\leq \frac{\Delta N}{N(\Delta N - 1)} \left[N - D_n(cq) + c(qN)N\left(\frac{\pi(\log\lfloor\frac{M}{4}\rfloor + 1)}{\Delta N} + \frac{\pi^2}{6(\Delta N)^2}\right)\right]^2.\end{aligned}$$

iv) We use the estimates for the Dirichlet kernel $N - D_n(q) \geq (qN)^2 N$ in ii) and $N - D_n(cq) \leq N\frac{\pi^2}{6}c^2(qN)^2$ in iii), see Lemma 2.4.7, and insert this in (3.3.12) to get finally

$$\begin{aligned}\left\|(\boldsymbol{K}_2 - \boldsymbol{B}\boldsymbol{K}_1^{-1}\boldsymbol{B}^*)^{-1}\right\| \leq \frac{1}{N(qN)^2}\Bigg[2 &- \frac{c^2\pi^2(\log\lfloor\frac{M}{4}\rfloor + 1)}{\Delta N} - \frac{c^2\pi^3}{3(\Delta N)^2} - \frac{2.42c^2}{(\Delta N)^3} \\ &- \frac{\Delta N}{\Delta N - 1}\left(\frac{c^2\pi^2}{6}(qN) + \frac{c\pi(\log\lfloor\frac{M}{4}\rfloor + 1)}{\Delta N} + \frac{c\pi^2}{6(\Delta N)^2}\right)^2\Bigg]^{-1}.\end{aligned}$$

This upper bound also bounds the maximum in (3.3.11) since for all $qN \leq \frac{1}{2}$ and $\Delta N \geq 2$ together with Theorem 3.1.7

$$\left\|\boldsymbol{K}_1^{-1}\right\| \leq \frac{2}{N} \leq \frac{1}{2N(qN)^2} \leq \frac{1}{N(qN)^2}[2 - \cdots]^{-1}.$$

□

Theorem 3.3.13 (Upper bound)**.**
Under the conditions of (3.3.9) *with* $M \geq 4$, $qN \leq q_{\max}N = \frac{1}{4c^2}$ *and* $\Delta N \geq \Delta_{\min}N = 10c^2(\log\lfloor\frac{M}{4}\rfloor + 1)$, *we have*

$$\operatorname{cond}(\boldsymbol{A}) \leq \frac{5}{qN}.$$

Proof. In Lemma 3.3.12 the constant $C(qN, \Delta N, c, M)$ is monotone increasing in qN and monotone decreasing in ΔN. Hence, after plugging in the bounds for ΔN and ΔN in our assumptions, it is easy to see that the constant $C(\frac{1}{4c^2}, 10c^2(\log\lfloor\frac{M}{4}\rfloor + 1), c, M)$ is monotone decreasing in c and M, respectively. Therefore, we get $C(qN, \Delta N, c, M) \leq C(1/4, 10, 1, 4) \leq 11.3$, so that $\left\|\boldsymbol{K}^{-1}\right\| \leq 11.3N^{-1}(qN)^{-2}$. Together with the bound $\|\boldsymbol{K}\| \leq 22N/10 = 2.2N$ from Lemma 3.3.9, we obtain the result. □

If each pair in the clusters has the same separation distance, i.e. $c = 1$ in (3.3.9), we can improve the upper bound in the sense that restrictions on q except for $qN \leq 1$ can be dropped. In order to obtain the same constant, we have to increase the restrictions on Δ slightly.

Lemma 3.3.14.
Under the conditions of (3.3.9) *with* $c = 1$*, such that*

$$\begin{aligned}\widetilde{C}(\Delta N, M) := 2 &- \frac{\pi^2(\log\lfloor\frac{M}{4}\rfloor + 1)}{\Delta N} - \frac{\pi^3}{3(\Delta N)^2} - \frac{2.42}{(\Delta N)^3} \\ &- \frac{\Delta N}{\Delta N - 1} - \frac{2\pi(\log\lfloor\frac{M}{4}\rfloor + 1)}{\Delta N - 1} - \frac{\pi^2}{3(\Delta N)(\Delta N - 1)} \\ &- \frac{\pi^2(\log\lfloor\frac{M}{4}\rfloor + 1)^2}{(\Delta N)(\Delta N - 1)} - \frac{\pi^3(\log\lfloor\frac{M}{4}\rfloor + 1)}{3(\Delta N)^2(\Delta N - 1)} - \frac{\pi^4}{36(\Delta N)^3(\Delta N - 1)}\end{aligned}$$

is positive, we have

$$\left\|\boldsymbol{K}^{-1}\right\| \leq \frac{C(\Delta N, M)}{N(qN)^2},$$

where $C(\Delta N, M) := \left(\frac{2\Delta N}{\Delta N - 1} + \sqrt{\frac{\Delta N + 1}{\Delta N - 1}}\right) / \widetilde{C}(\Delta N, M)$.

Proof. The proof is analogous to that of Lemma 3.3.12, the only difference is in step iv). Setting $c = 1$ in ii) and iii), expanding the squared bracket in iii) and inserting this into (3.3.12) leads to

$$\begin{aligned}\left\|(\boldsymbol{K}_2 - \boldsymbol{B}\boldsymbol{K}_1^{-1}\boldsymbol{B}^*)^{-1}\right\| \leq \Bigg[&2\left(N - D_n\left(q\right)\right) \\ &- (qN)^2 N\left(\frac{\pi^2(\log\lfloor\frac{M}{4}\rfloor + 1)}{\Delta N} + \frac{\pi^3}{3(\Delta N)^2} + \frac{2.42}{(\Delta N)^3}\right) - \frac{\Delta N}{N(\Delta N - 1)}\left(N - D_n\left(q\right)\right)^2 \\ &- \frac{\Delta N}{\Delta N - 1} 2qN\left(N - D_n\left(q\right)\right)\left(\frac{\pi(\log\lfloor\frac{M}{4}\rfloor + 1)}{\Delta N} + \frac{\pi^2}{6(\Delta N)^2}\right) \\ &- (qN)^2 N\frac{\Delta N}{\Delta N - 1}\left(\frac{\pi^2(\log\lfloor\frac{M}{4}\rfloor + 1)^2}{(\Delta N)^2} + \frac{\pi^3(\log\lfloor\frac{M}{4}\rfloor + 1)}{3(\Delta N)^3} + \frac{\pi^4}{36(\Delta N)^4}\right)\Bigg]^{-1}.\end{aligned}$$

In three summands, we can factor out $N - D_n\left(q\right)$ and use the estimate $N - D_n\left(q\right) \geq (qN)^2 N$, leading to a larger bound after inverting the expression in the end. Afterwards, in the third summand $N - D_n\left(q\right)$ is left, for which we use the rough bound $N - D_n\left(q\right) \leq N$. In the fourth summand we use $qN \leq 1$ for the single qN. The same argument as in (3.3.8) shows that this also bounds the maximum in (3.3.11) and we get the result. □

Theorem 3.3.15 (Upper bound)**.**
Under the conditions of (3.3.9) *with* $c = 1$, $\Delta N \geq \Delta_{\min} N = 25(\log\lfloor \frac{M}{4} \rfloor + 1)$, *we have*

$$\operatorname{cond}(\boldsymbol{A}) < \frac{5}{qN}.$$

Proof. Direct inspection gives monotonicity of $C(\Delta N, M)$ with respect to ΔN and also the estimate $C(25(\log\lfloor M/4 \rfloor + 1), M) \leq C(25, 4) \leq 12$. Hence $\left\|\boldsymbol{K}^{-1}\right\| \leq 12N^{-1}(qN)^{-2}$ and together with the bound $\|\boldsymbol{K}\| \leq 52N/25$ from Lemma 3.3.9 we obtain the result. □

Remark 3.3.16.
Due to Lemma 2.1.6, the upper bound from Theorem 3.3.13 remains valid if nodes are removed. We note in passing that $\sigma_{\min}$ *and* $\sigma_{\max}$ *are monotone increasing with* N *and thus, condition number estimates for an even number* N *follow. Lower and upper bounds in Lemma 3.3.1 and Theorem 3.3.13 finally yield*

$$\frac{1}{qN} \leq \operatorname{cond}(\boldsymbol{A}) \leq \frac{5}{qN}\,.$$

The lower bound is tight and the numerical value 5 *in the upper bound follows from our proof technique and can be improved, see Figure 3.3.6. The uniformity condition* $qN \leq 1/(4c^2)$ *is artificial and except for the special cases in Theorem 3.3.6 and Theorem 3.3.15, prevents letting* $\Delta \to 1/N$.

3.3.3 Numerical examples

All computations were carried out using MATLAB R2020a. As a test for the bounds in the case of one pair cluster, we use the following configuration. Let the number of nodes $M = 20$ and $M = 200$ be fixed, respectively. Moreover, we choose $N = 1 + 12(M - 1)$ which ensures that all nodes fit on the unit interval. We choose $q \in \left[10^{-11}/N, 1/N\right]$ logarithmically uniformly at random and $\Delta_3, \ldots, \Delta_M \in [6/N, 12/N]$ uniformly at random. Then, we set the nodes $t_1 < \cdots < t_M \in [0, 1)$ such that $t_1 = 0$, $t_2 = q$ and for $j = 3, \ldots, M$, $|t_j - t_{j-1}| = \Delta_j$. Afterwards, the condition number of the corresponding Vandermonde matrix is computed. This procedure is repeated 100 times and the results are presented in Figure 3.3.6 (left).

For pair clusters, we use the following configuration. Let the number of nodes $M = 20$ and $M = 200$ be fixed, respectively. Moreover, we choose the parameter $c = 2$ and q_{max} and $\Delta_{\min}$ as in Theorem 3.3.13. To ensure that all nodes fit on the unit interval, we choose N as the smallest odd integer bigger than $(cq_{\max}N + 2\Delta_{\min}N)M/2$. Then, we choose $q \in \left[10^{-11}/N, 1/N\right]$ logarithmically uniformly at random and set the nodes $t_1 < \cdots < t_M \in [0, 1)$ such that $t_1 = 0$, $t_2 = q$ and for $j = 3, \ldots, M$, $|t_j - t_{j-1}| = \Delta_j$ if j is odd or $|t_j - t_{j-1}| = q_j$ if j is even, where $q_j \in [q, cq]$ and $\Delta_j \in [\Delta_{\min}, 2\Delta_{\min}]$ are picked uniformly at random, respectively. Afterwards, the condition number of the corresponding Vandermonde matrix is computed. This procedure is repeated 100 times and the results are presented in Figure 3.3.6 (right). Note that Theorem 3.3.13 makes the restriction $q \leq q_{\max} = \frac{1}{4N}$, which seems to be an artifact of our proof technique.

3.3.4 Results independent of the number of nodes

During the peer review process of [68] one anonymous reviewer suggested, instead of analyzing the eigenvalues of $\check{\boldsymbol{A}}\check{\boldsymbol{A}}^*$ ($\check{\boldsymbol{A}} := \operatorname{diag}(z_1^{-n}, \ldots, z_M^{-n})\boldsymbol{A}$ denotes the shifted Vandermonde matrix)

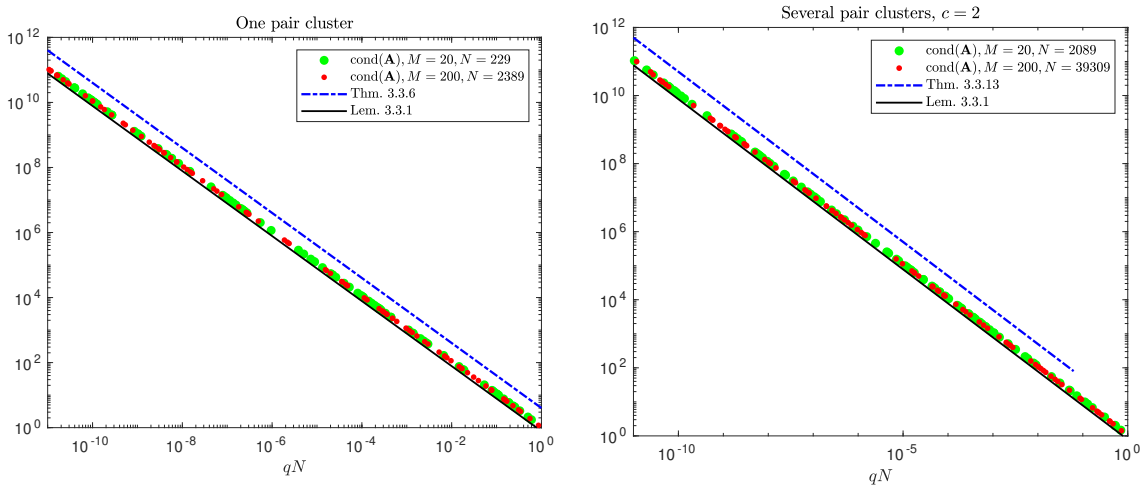


Figure 3.3.6: numerical experiments for bounds on the condition number, lower bounds from Lemma 3.3.1; left: one pair cluster, upper bound from Theorem 3.3.6; right: several pair clusters, upper bound from Theorem 3.3.13.

one could analyze the eigenvalues of $\check{\boldsymbol{A}}\boldsymbol{C}\check{\boldsymbol{A}}^*$ with appropriate diagonal matrix $\boldsymbol{C}$. For example choosing $\boldsymbol{C} := \operatorname{diag}(|k|/(n+1))_{|k|\leq n}$ leads to the kernel matrix $\check{\boldsymbol{A}}\boldsymbol{C}\check{\boldsymbol{A}}^*$ with entries

$$\left(\check{\boldsymbol{A}}\boldsymbol{C}\check{\boldsymbol{A}}^*\right)_{j,k} = F_n(t_j - t_k),$$

where $F_n \colon \mathbb{T} \to \mathbb{R}$,

$$F_n(t) := \sum_{k=-n}^{n} \left(1 - \frac{|k|}{n+1}\right) \mathrm{e}^{2\pi i k t}.$$

is the scaled well-known *Fejér kernel* of degree n. The Fejér kernel has quadratic decay away from zero since it is a squared Dirichlet kernel (cf. [89, p. 22] and see (A.0.1)) which has linear decay by Lemma 2.4.7. One can compare this to the second power of the modified Dirichlet kernel from Lemma 2.4.11 which is in absolute value a scaled Dirichlet kernel. In the lower bounds for the smallest singular value from the above section, this leads to the vanishing of logarithmic terms in the number of nodes M, i.e. bounds are independent of M in the end. A precondition for this, of course, is that $\lambda_{\min}(\check{\boldsymbol{A}}\boldsymbol{C}\check{\boldsymbol{A}}^*) \leq \lambda_{\min}(\check{\boldsymbol{A}}\check{\boldsymbol{A}}^*)$. That this is true can be simply seen, using Theorem 2.1.4 and $\|\boldsymbol{C}\| \leq 1$, by the calculation

$$\begin{aligned}\lambda_{\min}(\check{\boldsymbol{A}}\boldsymbol{C}\check{\boldsymbol{A}}^*) &= \min_{\boldsymbol{x}\in\mathbb{C}^M, \|\boldsymbol{x}\|=1} \boldsymbol{x}^*\check{\boldsymbol{A}}\boldsymbol{C}\check{\boldsymbol{A}}^*\boldsymbol{x} = \min_{\boldsymbol{x}\in\mathbb{C}^M, \|\boldsymbol{x}\|=1} \left\|\boldsymbol{C}^{1/2}\check{\boldsymbol{A}}^*\boldsymbol{x}\right\|^2 \\ &\leq \min_{\boldsymbol{x}\in\mathbb{C}^M, \|\boldsymbol{x}\|=1} \|\boldsymbol{C}\| \left\|\check{\boldsymbol{A}}^*\boldsymbol{x}\right\|^2 \leq \min_{\boldsymbol{x}\in\mathbb{C}^M, \|\boldsymbol{x}\|=1} \left\|\check{\boldsymbol{A}}^*\boldsymbol{x}\right\|^2 = \lambda_{\min}(\check{\boldsymbol{A}}\check{\boldsymbol{A}}^*).\end{aligned}$$

Alternatively, one can use Lemma 2.1.17 which moreover shows that $\check{\boldsymbol{A}}\boldsymbol{C}\check{\boldsymbol{A}}^*$ does not help for establishing an upper bound on the largest eigenvalue in the same way since $\lambda_{\min}(\check{\boldsymbol{A}}\boldsymbol{C}\check{\boldsymbol{A}}^*) \leq \lambda_{\min}(\check{\boldsymbol{A}}\check{\boldsymbol{A}}^*)$. It is also possible to see this in context of Lemma 3.1.6. Analogously to Theorems 3.3.6, 3.3.13 and 3.3.15 this approach yields

Theorem 3.3.17 (Upper bound, one pair cluster, cf. Theorem 3.3.6)**.**
Let $M, N = 2n+1 \in \mathbb{N}_+$, $M \leq N$ and $\boldsymbol{A} \in \mathbb{C}^{M\times N}$ as in (3.1.6)*. If the nodes satisfy the conditions from* (3.3.3) *with $\Delta N \geq \Delta_{\min} N = 6$, we have*

$$\operatorname{cond}(\boldsymbol{A}) \leq \frac{8.2}{qN}.$$

Theorem 3.3.18 (Upper bound, cf. Theorem 3.3.13)**.**
Let $M, N = 2n+1 \in \mathbb{N}_+$, $M \leq N$ and $\boldsymbol{A} \in \mathbb{C}^{M\times N}$ as in (3.1.6)*. If the nodes satisfy the conditions from* (3.3.9) *with $qN \leq q_{\max} N = \frac{1}{4c^2}$ and $\Delta N \geq \Delta_{\min} N = 11c^2$, we have*

$$\operatorname{cond}(\boldsymbol{A}) \leq \frac{14}{qN}.$$

Theorem 3.3.19 (Upper bound, uniform pair clusters, cf. Theorem 3.3.15)**.**
Let $M, N = 2n+1 \in \mathbb{N}_+$, $M \leq N$ and $\boldsymbol{A} \in \mathbb{C}^{M\times N}$ as in (3.1.6)*. If the nodes satisfy the conditions from* (3.3.9) *with $c = 1$, $\Delta N \geq \Delta_{\min} N = 25$, we have*

$$\operatorname{cond}(\boldsymbol{A}) < \frac{8.3}{qN}.$$

The detailed proofs are given in Appendix A as they are analogue to the ones presented in the previous sections.

3.3.5 Limitation of the technique

In principle, the suggested Schur-complement technique can be generalized to more than two nodes per cluster. Let $n, M \in \mathbb{N}_+$, $M \geq 3$, $N = 2n+1 > M$ and $0 = t_1 < \cdots < t_M \in \mathbb{T}$ be such that $\{t_1, t_2, t_3\}$ build a cluster and decompose

$$\boldsymbol{K} = \begin{pmatrix} \boldsymbol{K}_1 & \boldsymbol{B}^* \\ \boldsymbol{B} & \boldsymbol{K}_2 \end{pmatrix}, \quad \boldsymbol{K}_1 = \begin{pmatrix} N & D_n(t_1 - t_2) \\ D_n(t_1 - t_2) & N \end{pmatrix}, \quad \boldsymbol{K}_2 = \left(D_n(t_i - t_j)\right)_{i,j=3}^{M}.$$

While it is clear that the Schur-complement $\boldsymbol{K}_1 - \boldsymbol{B}^* \boldsymbol{K}_2^{-1} \boldsymbol{B}$ is strictly positive definite, establishing a lower bound on its smallest singular value similar to the proofs of Lemmata 3.3.4 and 3.3.5 seems considerably harder. Already the linear approximation in Lemma 3.3.4 then needs to be replaced by a higher order approximation for the matrix $\boldsymbol{B}$.

Also a multivariate extensions may possible albeit with highly technical effort. We present a short supporting example. Consider the bivariate case and the Vandermonde matrix

$$\boldsymbol{A} = \left(z_j^{\boldsymbol{\nu}}\right)_{\substack{j=1,\dots,M \\ \boldsymbol{\nu}\in\mathbb{Z}^2,\, \|\boldsymbol{\nu}\|_\infty \leq n}} \in \mathbb{C}^{M\times N^2},$$

where $\boldsymbol{z}_j = (x_j, y_j) = (\mathrm{e}^{-2\pi i u_j}, \mathrm{e}^{-2\pi i v_j}) \in \mathbb{T}^2$, $\boldsymbol{\nu} = (\alpha, \beta) \in \mathbb{Z}^2$ is a multi-index, and $\boldsymbol{z}_j^{\boldsymbol{\nu}} = x_j^\alpha \cdot y_j^\beta$. The distance of the nodes $\boldsymbol{t}_j = (u_j, v_j) \in [0,1)^2$ is measured by $|\boldsymbol{t}_j - \boldsymbol{t}_\ell|_{\mathbb{T}} := \min_{\boldsymbol{r}\in\mathbb{Z}^2} \|\boldsymbol{t}_j - \boldsymbol{t}_\ell + \boldsymbol{r}\|_\infty$ and we consider the situation as in (3.3.3) and Definition 3.3.2 with $\boldsymbol{K} = \boldsymbol{A}\boldsymbol{A}^*$. Lemma 3.3.4 can be proven using the bivariate mean value theorem to get $|r_j| \leq (qN)N\pi / |\xi_j|_{\mathbb{T}}$, $j = 2, 3, \dots, M$, and the packing argument [71, Lem. 4.5] to get

$$\|\boldsymbol{r}\|^2 \leq (N^2 - D_n(u_2)D_n(v_2))^2 + \frac{12\pi^2 N^4 (qN)^2}{(\Delta N - 1)^2} \left(1 + \log\left\lceil \sqrt{M/6} \right\rceil\right).$$

We need additional assumptions for Lemma 3.3.5 to work since results for general well separated nodes, cf. [66], seem to be too weak. If the nodes $t_2, \dots, t_M$ are a subset of equispaced nodes in $\mathbb{T}^2$, then [71, Cor. 4.11] yields $\left\|\boldsymbol{K}_2^{-1}\right\| \leq (N - 1/\Delta))^{-2}$. Together with $M \geq 4$ and $\Delta N \geq 4 + 2\log(M)$, this yields $\left\|\boldsymbol{K}^{-1}\right\| \leq 20/N^2(qN)^2$.

We see in the next section that the technique introduced by Li and Liao [73] for the univariate case is better qualified for dealing with multiple nodes in the clusters and we show, that it is extendable to arbitrary dimensions.

3.4 Larger clusters and multivariate extension

We confine the subsequent analysis to multivariate Vandermonde matrices with monomial degrees up to a certain max-degree that naturally arises when taking the tensor product of univariate polynomials up to a certain degree.

Let $\boldsymbol{\nu} := (\nu_1, \dots, \nu_d)^\top \in \mathbb{N}^d$ be a multi-index, $n \in \mathbb{N}$ be a degree and $N := n + 1$ and assume $M \in \mathbb{N}_+$, $M \leq N^d$. Again we have the a node set $\Omega = \{\boldsymbol{t}_1, \dots, \boldsymbol{t}_M\} \subset \mathbb{T}^d$ and $\boldsymbol{z}_j = (\mathrm{e}^{2\pi i(\boldsymbol{t})_1}, \dots, \mathrm{e}^{2\pi i(\boldsymbol{t})_d})^\top \in \left\{\boldsymbol{z} \in \mathbb{C}^d \mid |(\boldsymbol{z})_j| = 1\right\}$, $j = 1, \dots, M$. We are interested in the multivariate, rectangular Vandermonde matrix

$$\boldsymbol{A} := \boldsymbol{A}_N(\Omega) := \left(\boldsymbol{z}_j^{\boldsymbol{\nu}}\right)_{\substack{j=1,\dots,M \\ \boldsymbol{\nu}\in\mathbb{N}^d,\, \|\boldsymbol{\nu}\|_\infty < N}} \in \mathbb{C}^{M \times N^d}, \qquad \boldsymbol{z}_j^{\boldsymbol{\nu}} := (\boldsymbol{z}_j)_1^{\nu_1} \cdot \dots \cdot (\boldsymbol{z}_j)_d^{\nu_d}, \tag{3.4.1}$$

of degree $N - 1$. Similar to the univariate case we utilize the following distances.

Definition 3.4.1 (Distances)**.**
The wrap-around distance *between two nodes $\boldsymbol{t}, \boldsymbol{t}' \in \mathbb{T}^d$ is defined by*

$$\left|\boldsymbol{t} - \boldsymbol{t}'\right|_{\mathbb{T}^d} := \min_{\boldsymbol{r}\in\mathbb{Z}^d} \left\|\boldsymbol{t} - \boldsymbol{t}' + \boldsymbol{r}\right\|_\infty .$$

We note that this distance is the largest wrap-around distance in the coordinate directions. The minimal separation distance *of a node set $\Omega \subset \mathbb{T}^d$ is given by*

$$q := \min_{\boldsymbol{t}\neq\boldsymbol{t}'} \left|\boldsymbol{t} - \boldsymbol{t}'\right|_{\mathbb{T}^d} .$$

Taking the same symbol we used in the univariate case is justified since for, $d = 1$, this is the univariate wrap-around distance.

In one dimension, we used at some points (in (3.3.6) and the proofs of Lemmata 3.3.10 and 3.3.11) a packing argument to bound sums where distances from one node to all others appeared. There, relative to a node at most two nodes could appear, one on each side of the interval, not closer than the minimal separation distance, respectively. The next nodes have at least double the minimal separation distance and so on. The same principle we find more complicated in the multivariate setting. Here the space around nodes is partitioned into shells and each shell has room for at most a certain number of nodes which is made precise in the following lemma.

Lemma 3.4.2 (Partitioning into shells, [71, Def. 4.4, Lem. 4.5])**.**
For $d \in \mathbb{N}_+$ and a separation distance $0 < q \leq \frac{1}{2}$ the Partitioning of $\mathbb{T}^d$ into shells *is give by*

$$J_m := J_m(q) := \left\{\boldsymbol{t} \in \mathbb{T}^d \colon mq \leq |\boldsymbol{t}|_{\mathbb{T}^d} < (m+1)q\right\}, \qquad m = 0, \dots, \left\lfloor \frac{1}{2q} \right\rfloor - 1$$

and

$$J_{\left\lfloor \frac{1}{2q} \right\rfloor} := J_{\left\lfloor \frac{1}{2q} \right\rfloor}(q) := \left\{ \boldsymbol{t} \in \mathbb{T}^d \colon \left\lfloor \frac{1}{2q} \right\rfloor q \leq |\boldsymbol{t}|_{\mathbb{T}^d} \leq \frac{1}{2} \right\}.$$

If $\Omega \in \mathbb{T}^d$ is a node set, then $J_m(q) \cap \Omega$ has cardinality

$$|J_m(q) \cap \Omega| \leq 2^d(2^d - 1)m^{d-1}, \quad m = 1, \ldots, \left\lfloor \frac{1}{2q} \right\rfloor.$$

Proof. For completeness, we include the short proof that relies on a packing argument. Each shell $J_m, m \geq 1$, can be subdivided into shifted and rotated versions of the cube $[0, q)^d$, cf. Figure 3.4.1. This is done in a way that each point of J_m is contained in at least one of these cubes and the cubes do not share interior points. Now, each cube contains at most one node and therefore, we can estimate with the binomial theorem

$$\begin{aligned} |J_m(q) \cap \Omega| &\leq \frac{1}{q^d} \int_{J_m} 1 \, d\boldsymbol{x} \leq 2^d \left((m+1)^d - m^d \right) = 2^d \left(\sum_{\ell=0}^{d} \binom{d}{\ell} m^{d-\ell} \; - m^d \right) \\ &= 2^d \sum_{\ell=1}^{d} \binom{d}{\ell} m^{d-\ell} \leq 2^d m^{d-1} \sum_{\ell=1}^{d} \binom{d}{\ell} = 2^d(2^d - 1)m^{d-1}. \end{aligned}$$

□

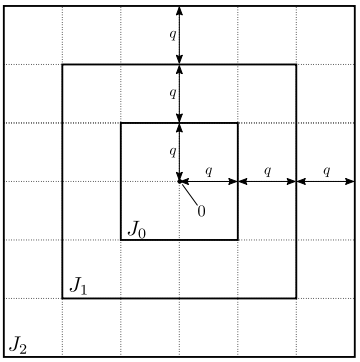


Figure 3.4.1: partitioning of $\mathbb{T}^2$ into shells. First 3 shells with distance q.

3.4.1 Multivariate clustered node configurations

To the best of our knowledge, the situation of clustered node configurations in multiple dimensions has not been dealt with in the literature so far. Most of the following results concerning the smallest singular value of multivariate Vandermonde matrices are published in [69].

We work out a multivariate extension of Theorem 2.3 in [73] and meanwhile improve it in a way that it becomes superior even applied to the univariate case, i.e. we do an improvement on the cluster separation condition, especially we make it independent of the number of nodes M. Furthermore, we provide an improved estimate on the smallest singular value $\sigma_{\min}(\boldsymbol{A})$ only depending on the biggest cluster size λ and not on the total number of nodes M. We

also provide bounds for the largest singular value of the multivariate Vandermonde matrix that uses the additional geometric information of the node set.

The following definition is the multivariate version of Definition 3.2.2 with an extension by the cluster complexity.

Definition 3.4.3 (Multivariate clustered node configurations)**.**
Let Ω be the node set of a Vandermonde matrix of degree $N-1$ given by (3.4.1)*.*

i) A subset of is called cluster *if it is contained in a cube of length $1/N$. For two clusters $\Lambda', \Lambda'' \subset \Omega$, we define*

$$\operatorname{dist}(\Lambda', \Lambda'') := \min\{\left|\boldsymbol{t}' - \boldsymbol{t}''\right|_{\mathbb{T}^d} : \ \boldsymbol{t}' \in \Lambda', \ \boldsymbol{t}'' \in \Lambda''\}.$$

ii) The node set Ω is called a clustered node configuration *with L clusters if it can be written as*

$$\Omega = \bigcup_{l=1}^{S} \Lambda_l,$$

where the Λ_l are clusters and the minimal cluster separation Δ *fulfills*

$$\Delta := \min_{1 \le l < l' \le S} \operatorname{dist}(\Lambda_l, \Lambda_{l'}) > \frac{1}{N}.$$

We order $|\Lambda_1| \ge |\Lambda_2| \ge \cdots \ge |\Lambda_L|$ and denote the cardinality of the biggest cluster by $\lambda := |\Lambda_1|$. In passing, the node set Ω is called well separated *with normalized separation ρ if $\lambda = 1$.*

iii) The cluster complexity *is defined by*

$$\mathcal{C} := \mathcal{C}(\Omega, N) := \max_{j=1,\dots,M} \prod_{\boldsymbol{t}' \in \Omega \colon 0 < |\boldsymbol{t}_j - \boldsymbol{t}'|_{\mathbb{T}^d} \le 1/N} \frac{1}{N \left|\boldsymbol{t}_j - \boldsymbol{t}'\right|_{\mathbb{T}^d}}$$

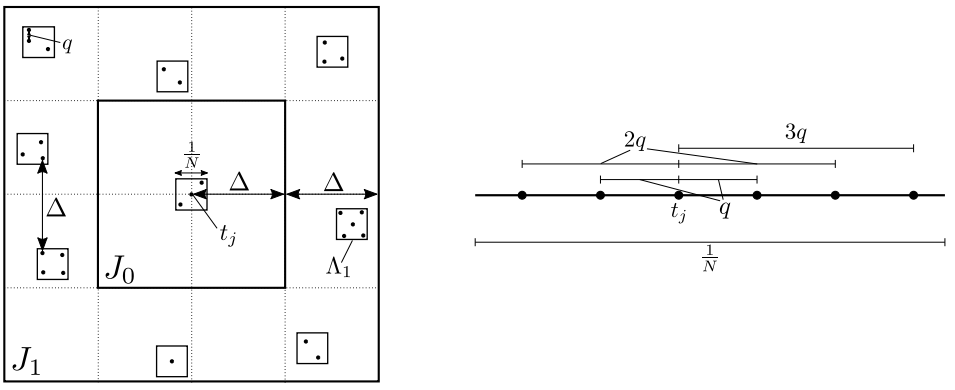


Figure 3.4.2: left: clustered node configuration and partitioning into shells for $d = 2$; right: a cluster which maximizes $\mathcal{C}$ for $d = 1$.

Remark 3.4.4 (Geometry of nodes and conditions on the degree)**.**
With the notation of Definition 3.4.3, we note that

i) the inequality $\sin(x) \geq 2x/\pi$ *for* $0 \leq x \leq \pi/2$ *implies*

$$\left|z - z'\right| = 2\sin(\pi \left|t - t'\right|_{\mathbb{T}}) \geq 4\left|t - t'\right|_{\mathbb{T}}, \qquad z := \mathrm{e}^{2\pi i t},\ z' := \mathrm{e}^{2\pi i t'} \in \mathbb{T}. \tag{3.4.2}$$

A higher order approximation is given in [73],

$$\left|z - z'\right| \geq 2\pi \left(1 - \frac{\pi^2 \left|t - t'\right|_{\mathbb{T}}^2}{3}\right)^{1/2} \left|t - t'\right|_{\mathbb{T}}. \tag{3.4.3}$$

ii) A necessary condition for the existence of a clustered node configuration with S *clusters is* $S^{1/d}\Delta \leq 1$*, with equality if and only if all nodes are equispaced. Similarly, if* $1 > S^{1/d}(\Delta + \frac{1}{N})$ *or equivalently*

$$N > \frac{S^{1/d}}{1 - \Delta S^{1/d}}$$

as condition for N*, then equispaced cluster with arbitrary node configuration within each cluster exist. Moreover, if assumptions on the clustered node configuration are only done via a minimal cluster separation, when lower bounds for the smallest singular value are established, then* Δ *needs to be at least linearly dependent on the biggest cluster size* λ*. If on the contrary,* $\Delta N < \lambda/4$ *and* $d = 1$ *for simplicity of the argument, then let* λ *nodes form a cluster of length* $1/N$ *and place one node as far as possible away. With fixed* N*, we have* $2\Delta N = N - 1$ *and therefore,* $\Delta N < \lambda/4$ *is equivalent to* $N \leq \lambda/2 + 1/2 < M/2 + 1/2$ *and thus* $\operatorname{rank}(\boldsymbol{A}) \leq N < M$*. On the other hand,* $\Delta N > \lambda$ *already implies* $N \geq S(\Delta N + 1) \geq S\lambda \geq M$*.*

The cluster complexity can be upper bounded by the minimal separation as described in the following lemma.

Lemma 3.4.5 (Bounds on the cluster complexity)**.**
Let $d \in \mathbb{N}_+$ *and for a clustered node configuration* $\Omega \in \mathbb{T}^d$ *let the cluster complexity* $\mathcal{C}$ *be as in Definition 3.4.3 iii). Then we have* $\mathcal{C} \leq (qN)^{1-\lambda}$ *and equality for* $\lambda = 1$ *and* $\lambda = 2$*. Refined for* $d = 1$*, it is easy to see that the cluster complexity is maximized by an equispaced cluster with* λ *nodes separated by* q *and taking distances from the center node, see Figure 3.4.2 (right). We thus have*

$$\mathcal{C} \leq \frac{1}{(qN)^{\lambda-1}} \left(\left\lfloor \frac{\lambda-1}{2} \right\rfloor ! \cdot \left\lceil \frac{\lambda-1}{2} \right\rceil !\right)^{-1} \leq \frac{1}{(qN)^{\lambda-1}\Gamma\left(\frac{\lambda+1}{2}\right)^2} \leq \frac{(2\mathrm{e})^{\lambda-1}}{\lambda^\lambda} \cdot \frac{1}{(qN)^{\lambda-1}} \tag{3.4.4}$$

and

$$\max_{\Omega} \mathcal{C} = \frac{1}{(qN)^{\lambda-1}} \left(\left\lfloor \frac{\lambda-1}{2} \right\rfloor ! \cdot \left\lceil \frac{\lambda-1}{2} \right\rceil !\right)^{-1} \geq \frac{(2\mathrm{e})^{\lambda-1}}{\lambda^{\lambda+1}} \cdot \frac{1}{(qN)^{\lambda-1}}, \tag{3.4.5}$$

where the maximum is taken over all clustered node configurations with minimal separation q *and the largest cluster containing* λ *nodes.*

Proof. The bounds in (3.4.4) and (3.4.5) follow by properties of the Gamma function, its logarithmic convexity, direct calculation, and Stirling's approximation from Lemma 2.4.1. □

3.4.2 The largest singular value

First of all, we look at an upper bound on the largest singular value. The trivial bound

$$\sigma_{\max}(\boldsymbol{A}) \leq \sqrt{N^d M} \tag{3.4.6}$$

always holds because each entry of $\boldsymbol{A}$ fulfills $\left|(\boldsymbol{A})_{j,\boldsymbol{\nu}}\right| = 1$ and by Lemma 2.1.8 we have $\sigma_{\min}(\boldsymbol{A}) = \|\boldsymbol{A}\| \leq \|\boldsymbol{A}\|_{\mathrm{F}}$. In the situation when the number of nodes is large this may not be a satisfying bound. The bound for the univariate case in Theorem 3.1.7 for example is valid for all node sets of distinct nodes, independent of M. Though, in the current situation with clustered nodes its dependency on the minimal separation distance is also disadvantageous. The next lemma shows that using the additional geometric information coming along with a clustered node configuration can be used to establish more accurate upper bounds on the largest singular value of the corresponding Vandermonde matrix. It relies on the same idea used in Section 3.3, namely splitting the node set into well separated subsets. The result becomes most effective in the situation where the number of nodes in the largest cluster λ is small compared to the total number of nodes M or when the cluster separation is relatively large.

Lemma 3.4.6 (Bounds on the largest singular value)**.**
Let the dimension $d \in \mathbb{N}_+$ and $\boldsymbol{A} \in \mathbb{C}^{M \times N^d}, M, N \in \mathbb{N}_+$ be a Vandermonde matrix corresponding to a clustered node configuration Ω with largest cluster size λ and cluster separation Δ, then

$$\sqrt{N^d \lambda} \leq \sup_{\Omega} \sigma_{\max}(\boldsymbol{A}) \leq \sqrt{N^d \lambda \left(1 + \frac{1}{\Delta N}\right)^d} \leq \sqrt{2^d N^d \lambda},$$

where the supremum is taken over all node sets satisfying the above properties. Furthermore, the upper bounds are smaller than the trivial bound $\sqrt{N^d M}$ if $\lambda < M$ and $\Delta N \geq \frac{\lambda^{1/d}}{M^{1/d} - \lambda^{1/d}}$ in the first case or $\lambda \leq \frac{M}{2^d}$ in the second case.

Proof. For the lower bound we use Remark 3.1.12 and consider the matrix $\widetilde{\boldsymbol{A}}$ with the subset $\widetilde{\Omega} = \{\boldsymbol{t}_1, \ldots, \boldsymbol{t}_\lambda\} \subset \Omega$, the nodes from a largest cluster. Let x be the largest separation distance of these nodes, $h := \max_{\boldsymbol{t} \neq \boldsymbol{t}'} |\boldsymbol{t} - \boldsymbol{t}'|_{\mathbb{T}^d}$. Since there was no assumption made for how close the nodes can be inside the cluster, we can look at $\widetilde{\boldsymbol{A}}\widetilde{\boldsymbol{A}}^*$ for decreasing distances until, in the limit, the nodes meet in one point. We have

$$\sigma_{\max}(\boldsymbol{A}) \geq \sigma_{\max}(\widetilde{\boldsymbol{A}})$$

with Remark 3.1.12. Taking the limit $h \to 0$ yields $\widetilde{\boldsymbol{A}} = \mathbb{1}_{\lambda \times N^d}$, where $\mathbb{1}_{\lambda \times N^d} \in \mathbb{C}^{\lambda \times N^d}$ is the matrix of all entries equal to one. Therefore, we have

$$\sup_{\Omega} \sigma_{\max}(\boldsymbol{A}) \geq \lim_{h \to 0} \sigma_{\min}(\widetilde{\boldsymbol{A}}) \geq \left\|\mathbb{1}_{\lambda \times N^d}\right\| = \sqrt{N^d \lambda}.$$

The last equality holds since $\left\|\mathbb{1}_{\lambda \times N^d}\right\| \leq \left\|\mathbb{1}_{\lambda \times N^d}\right\|_{\mathrm{F}} = \sqrt{\lambda N^d}$ on the one hand and $\left\|\mathbb{1}_{\lambda \times N^d}\right\| \geq \sqrt{\lambda N^d}$ on the other hand by choosing $\boldsymbol{x} = (\frac{1}{\sqrt{N^d}}, \ldots, \frac{1}{\sqrt{N^d}})^\top \in \mathbb{C}^\lambda$ as test vector in the definition of the norm.

For the upper bounds we assume the clustered node configuration consists of S clusters with at most λ nodes each. We pick one node from each cluster and collect the corresponding

nodes in a matrix. This can be done λ times (possibly without picking nodes from smaller clusters at some point) and we obtain submatrices $\boldsymbol{A}_\ell \in \mathbb{C}^{r_\ell \times N}, r_\ell \leq S, \ell = 1, \ldots, \lambda$ which contain the rows of $\boldsymbol{A}$ corresponding to the picked nodes. Since, a permutation of the rows does not affect the maximal singular value, we can reorder the nodes and assume that

$$\boldsymbol{A} = \begin{pmatrix} \boldsymbol{A}_1 \\ \vdots \\ \boldsymbol{A}_\lambda \end{pmatrix}.$$

Now, each matrix $\boldsymbol{A}_\ell$ is a Vandermonde matrix with well-separated nodes and minimal node separation distance at least $\Delta > 1/N$. Hence, by Lemma 3.5.11 (a forward reference to an independently proved result), which provides an upper bound on the largest singular value of multivariate Vandermonde matrices with well-separated nodes, we have for each $\ell = 1, \ldots, \lambda$

$$\|\boldsymbol{A}_\ell\|^2 \leq N^d \left(1 + \frac{1}{\Delta N}\right)^d.$$

Let $i(\ell, k), 1 \leq \ell \leq \lambda, 1 \leq k \leq r_\ell$ be the row index of $\boldsymbol{A}$, where we find the k-th row of $\boldsymbol{A}_\ell$. Applying Lemma 2.1.12, now yields

$$\begin{aligned}
\sigma_{\max}(\boldsymbol{A})^2 &= \max_{\boldsymbol{v} \in \mathbb{C}^N, \|\boldsymbol{v}\|=1} \|\boldsymbol{A}\boldsymbol{v}\|^2 = \max_{\boldsymbol{v} \in \mathbb{C}^N, \|\boldsymbol{v}\|=1} \sum_{j=1}^{M} \left|(\boldsymbol{A}\boldsymbol{v})_j\right|^2 \\
&= \max_{\boldsymbol{v} \in \mathbb{C}^N, \|\boldsymbol{v}\|=1} \sum_{\ell=1}^{\lambda} \sum_{k=1}^{r_\ell} \left|(\boldsymbol{A}\boldsymbol{v})_{i(\ell,k)}\right|^2 = \max_{\boldsymbol{v} \in \mathbb{C}^N, \|\boldsymbol{v}\|=1} \sum_{\ell=1}^{\lambda} \sum_{k=1}^{r_\ell} |(\boldsymbol{A}_\ell \boldsymbol{v})_k|^2 \\
&\leq \sum_{\ell=1}^{\lambda} \max_{\boldsymbol{v} \in \mathbb{C}^N, \|\boldsymbol{v}\|=1} \sum_{k=1}^{r_\ell} |(\boldsymbol{A}_\ell \boldsymbol{v})_k|^2 = \sum_{\ell=1}^{\lambda} \|\boldsymbol{A}_\ell\|^2 \leq \lambda N^d \left(1 + \frac{1}{\Delta N}\right)^d,
\end{aligned}$$

which provides the first inequality. Furthermore, $\Delta > 1/N$ and $1 + \frac{1}{\Delta N}$ is monotone decreasing in Δ. Therefore, we get the second and Δ-independent bound $\lambda N^d \left(\frac{\Delta N+1}{\Delta N}\right)^d \leq \lambda 2^d N^d$. Finally, if $\lambda < M$, then $\lambda \left(\frac{\Delta N+1}{\Delta N}\right)^d \leq M$ if and only if $\Delta N \geq \frac{\lambda^{1/d}}{M^{1/d} - \lambda^{1/d}}$, and $2^d \lambda \leq M$ if and only if $\lambda \leq \frac{M}{2^d}$. □

3.4.3 Lower bound on the smallest singular value

In the beginning of this chapter we have seen that Vandermonde matrices are closely related to polynomial interpolation since its entries are monomials evaluated at its nodes. Therefore, the following lemma connects the smallest singular value of the Vandermonde matrix to the L^2-norm of certain interpolating trigonometric polynomials. These polynomials live in the space specified by the monomials corresponding to the columns of the matrix.

Lemma 3.4.7 (Duality, cf. [73, Prop. 2.12])**.**
Let $d \in \mathbb{N}_+$ and $\boldsymbol{A} \in \mathbb{C}^{M \times N^d}, M \leq N$, a Vandermonde matrix with distinct nodes as in (3.4.1) *and $\boldsymbol{A}^\dagger$ its Moore-Penrose pseudo inverse from Definition 2.1.24. If $\sigma_{\max}(\boldsymbol{A}^\dagger) = \left\|\boldsymbol{A}^\dagger \boldsymbol{v}\right\|$ for some $\boldsymbol{v} = (v_1, \ldots, v_M)^\top \in \mathbb{C}^M, \|\boldsymbol{v}\| = 1$, then we have*

$$\sigma_{\min}(\boldsymbol{A}) = \max_{f \in \mathcal{P}(N-1),\, f(\boldsymbol{t}_j) = v_j, j=1,\ldots,M} \|f\|^{-1}_{L^2(\mathbb{T}^d)},$$

where $\mathcal{P}(N-1)$ is the space of trigonometric polynomials of max-degree at most $N-1$ from Definition 2.3.2.

Proof. By taking a Lagrange basis and using $M \leq N$ ensures that there exists a trigonometric polynomial interpolating the vector $\boldsymbol{v}$ at the nodes. By Parseval's identity, Theorem 2.3.8, we have $\|f\|_{L^2(\mathbb{T}^d)} = \left\|\widehat{\boldsymbol{f}}\right\|$, where $\widehat{\boldsymbol{f}}$ denotes the vector of the Fourier coefficients of f. Therefore, the interpolating function $f \in \mathcal{P}(N-1)$ having the least L^2-norm also has the least normed Fourier coefficients. The interpolation condition is equivalent to the linear system

$$\boldsymbol{A}\widehat{\boldsymbol{f}} = \boldsymbol{v}$$

that has solutions with least norm given by $\boldsymbol{A}^\dagger \boldsymbol{v}$, [58, Fact 5.8]. Finally, we have

$$\min_{f\in\mathcal{P}(N-1),\, f(\boldsymbol{t}_j)=v_j, j=1,\dots,M} \|f\|_{L^2(\mathbb{T}^d)} = \min_{\boldsymbol{u}\in\mathbb{C}^{N^d},\, \boldsymbol{A}\boldsymbol{u}=\boldsymbol{v}} \|\boldsymbol{u}\| = \left\|\boldsymbol{A}^\dagger \boldsymbol{v}\right\| = \left\|\boldsymbol{A}^\dagger\right\| = \frac{1}{\sigma_{\min}(\boldsymbol{A})}.$$

□

Although the preceding lemma allows a better understanding of the smallest singular value of Vandermonde matrices it is not easy to use it for bounding the latter. Especially, when nodes are clustering and the polynomial degree is limited, it is difficult to control the L^2-norm of interpolating function at these nodes. Additionally, finding interpolating polynomials is a challenging task on its own. In order to deal with these problems Li and Liao developed a so called robust duality lemma. It builds the core of the technique developed in [73] and we adapt it here to the multivariate setting. The advantage of that lemma is that we do not need to have an exact interpolating function. Instead if $\boldsymbol{v} \in \mathbb{C}^M$ is a unit norm vector such that $\sigma_{\min}(\boldsymbol{A}) = \|\boldsymbol{A}^*\boldsymbol{v}\|_2$, it suffices to construct a function $f \in \mathcal{P}(N-1)$ only almost interpolating the values of $\boldsymbol{v}$ in order to provide a lower bound. The task becomes also easier because a larger class of functions is considered.

Lemma 3.4.8 (Robust duality, cf. [73, Prop. 2.13])**.**
Let $M, N, d \in \mathbb{N}_+$. Let Ω and $\boldsymbol{A} \in \mathbb{C}^{M\times N^d}$ be given as in (3.4.1). If for a given unit norm vector $\boldsymbol{v} = (v_1, \dots, v_M)^\top \in \mathbb{C}^M$, and $\boldsymbol{\epsilon} = (\epsilon_1, \dots, \epsilon_M)^\top \in \mathbb{C}^M$ with $\|\boldsymbol{\epsilon}\|_2 \leq 1$, there exists a trigonometric polynomial f of max-degree at most $N-1 \in \mathbb{N}$, i.e., $f \in \mathcal{P}(N-1)$, see Definition 2.3.2, such that $f(\boldsymbol{t}_j) = v_j + \epsilon_j$ for each $j = 1, \dots, M$, then

$$\|\boldsymbol{A}^*\boldsymbol{v}\|_2 \geq (1 - \|\boldsymbol{\epsilon}\|_2) \|f\|^{-1}_{L^2(\mathbb{T}^d)}.$$

Proof. Define the discrete measure $\mu := \sum_{j=1}^M v_j \delta_{\boldsymbol{t}_j}$. Its Fourier coefficients are given by

$$\widehat{\mu}(\boldsymbol{\nu}) = \int_{\mathbb{T}^d} \mathrm{e}^{-2\pi i \boldsymbol{\nu}^* \boldsymbol{t}}\, d\mu(\boldsymbol{t}) = \sum_{j=1}^M v_j \boldsymbol{z}_j^{-\boldsymbol{\nu}} = (\boldsymbol{A}^*\boldsymbol{v})_{\boldsymbol{\nu}}, \quad \boldsymbol{\nu} \in \mathbb{N}^d, \|\boldsymbol{\nu}\|_\infty < N.$$

On the one hand, using the interpolation property of f and the lower triangular inequality of the absolute value, we have

$$\left|\int_{\mathbb{T}^d} \overline{f}\, d\mu\right| = \left|\sum_{j=1}^M \overline{f(\boldsymbol{t}_j)} v_j\right| = \left|\|\boldsymbol{v}\|_2^2 + \sum_{j=1}^M \overline{\epsilon_j} v_j\right| \geq \|\boldsymbol{v}\|_2^2 - \|\boldsymbol{v}\|_2 \|\boldsymbol{\epsilon}\|_2 = (1 - \|\boldsymbol{\epsilon}\|_2),$$

and on the other hand, using $f \in \mathcal{P}(N-1)$, the Cauchy–Schwarz inequality and Parseval's identity, we have

$$\left|\int_{\mathbb{T}^d} \overline{f}\, d\mu\right| = \left|\sum_{\boldsymbol{\nu}\in\mathbb{N}^d, \|\boldsymbol{\nu}\|_\infty \le N-1} \overline{\widehat{f}_{\boldsymbol{\nu}}}\widehat{\mu}(\boldsymbol{\nu})\right| \le \left\|\widehat{\boldsymbol{f}}\right\|_2 \|\boldsymbol{A}^*\boldsymbol{v}\|_2 = \|f\|_{L^2(\mathbb{T}^d)} \|\boldsymbol{A}^*\boldsymbol{v}\|_2 .$$

□

If $\boldsymbol{v}$ is chosen in Lemma 3.4.8 such that $\sigma_{\min}(\boldsymbol{A}) = \|\boldsymbol{A}^*\boldsymbol{v}\|$ a lower bound on the smallest singular value of $\boldsymbol{A}$ can be found by constructing an almost-interpolating function. Moreover, there are two ingredients that are important for obtaining a good (as large as possible) bound. On the one hand the interpolation error has to be small and on the other hand the function needs to have small L^2-norm.

Remark 3.4.9.
In the previous lemma, a bound on the smallest singular value on the Vandermonde matrix $\boldsymbol{A}$ is established in terms of the L^2-norm of an interpolating function. Notice that this function is living in the space of trigonometric polynomials determined by the set of monomials corresponding to the column index set of the Vandermonde matrix. Therefore, this lemma also holds true for Vandermonde matrices with columns indexed differently with respect to other function space, for instance the set of trigonometric polynomials with total degree less than N or hyperbolic cross polynomials, see e.g. [59].

The idea for constructing a function that allows to use the robust duality in Lemma 3.4.8 for clustered node configurations is the following. We construct Lagrange-like basis functions, where each of them belongs to a specific node. Each function has value one at its node, zero at nodes belonging to the same cluster and it rapidly decays outside the cluster to ensure that it is small at the remaining nodes in the set. The latter is the difference to the Lagrange basis and the reason for ending up with a function that is only almost-interpolating. A Lagrange-like basis function is constructed by multiplying a Lagrange basis function for nodes inside the cluster with a modified Dirichlet kernel raised to some power from Definition 2.4.10.

Lemma 3.4.10 (Lagrange-like basis with decay, cf. [73, Lem. A.1])**.**
Let $\beta, d, M, n, \lambda \in \mathbb{N}_+$, $\lambda \le M$, $N = n+1$, β be even, $\Omega = \{\boldsymbol{t}_1, \dots, \boldsymbol{t}_M\} \subset [0,1)^d$ be a clustered node configuration from Definition 3.4.3 and $n \ge 2\beta^2\lambda$. Then for each $\boldsymbol{t}_j \in \Omega$ with $\boldsymbol{t}_j \in \Lambda_l$ for some $l = l(j)$, there exists an $I_j \in \mathcal{P}(n)$, such that

i) $I_j(\boldsymbol{t}_k) = \delta_{jk}$ for all $\boldsymbol{t}_k \in \Lambda_l$, where $\delta_{j,k}$ denotes the Kronecker delta,

ii) $|I_j(\boldsymbol{t})| \le \frac{\beta^\beta \lambda^{\beta+\lambda-1}}{(2N|\boldsymbol{t}-\boldsymbol{t}_j|_{\mathbb{T}^d})^\beta}\mathcal{C}$ for all $\boldsymbol{t} \ne \boldsymbol{t}_j$, and

iii) $$\left|\langle I_k, I_j\rangle_{L^2(\mathbb{T}^d)}\right| \le \frac{\lambda^d\beta^{d/2}}{N^d}\lambda^{2\lambda-2}\mathcal{C}^2 \begin{cases} 1, & \boldsymbol{t}_k \in \Lambda_l, \\ \frac{\sqrt{\beta}}{2}\left(\frac{\lambda\beta}{N|\boldsymbol{t}_j-\boldsymbol{t}_k|_{\mathbb{T}^d}}\right)^\beta, & \text{otherwise.} \end{cases}$$

Proof. We define the functions I_j as product of a Lagrange polynomial G_j within the cluster and a fast decaying function H_j. Let $j \in \{1, \dots, M\}$ be fixed and define the j-th Lagrange

polynomial within its cluster Λ_l, $l = l(j)$, as follows. If $|\Lambda_l| = 1$, we simply set $G_j \equiv 1$. Otherwise, let

$$Q := \left\lfloor \frac{n}{\lambda} \right\rfloor \geq \frac{n-\lambda+1}{\lambda} \geq \frac{N}{2\lambda} \tag{3.4.7}$$

denote the "blow-up-factor" (that in principal will spread the cluster nodes) and for $\boldsymbol{t}_k \in \Lambda_l \setminus \{\boldsymbol{t}_j\}$ let $\ell(k)$ be the index of the vector component that realizes the distance $|\boldsymbol{t}_j - \boldsymbol{t}_k|_{\mathbb{T}^d}$. We immediately have $|Q\boldsymbol{t}_j - Q\boldsymbol{t}_k|_{\mathbb{T}^d} = Q\,|\boldsymbol{t}_j - \boldsymbol{t}_k|_{\mathbb{T}^d} \neq 0$ since nodes in the same cluster are at most $1/N$-separated and thus

$$G_j(\boldsymbol{t}) := \prod_{\boldsymbol{t}_k \in \Lambda_l \setminus \{\boldsymbol{t}_j\}} \frac{\mathrm{e}^{2\pi i Q(\boldsymbol{t})_{\ell(k)}} - \mathrm{e}^{2\pi i Q(\boldsymbol{t}_k)_{\ell(k)}}}{\mathrm{e}^{2\pi i Q(\boldsymbol{t}_j)_{\ell(k)}} - \mathrm{e}^{2\pi i Q(\boldsymbol{t}_k)_{\ell(k)}}}$$

fulfills $G_j(\boldsymbol{t}_k) = \delta_{j,k}$ and by the inequalities (3.4.2) and (3.4.7)

$$\|G_j\|_{\mathcal{C}(\mathbb{T}^d)} \leq \prod_{\boldsymbol{t}_k \in \Lambda_l \setminus \{\boldsymbol{t}_j\}} \frac{1}{2Q\left|(\boldsymbol{t}_j)_{\ell(k)} - (\boldsymbol{t}_k)_{\ell(k)}\right|_{\mathbb{T}}} \leq \lambda^{\lambda-1}\mathcal{C}. \tag{3.4.8}$$

We proceed by setting

$$P := \left\lfloor \frac{n}{\lambda\beta} \right\rfloor, \quad P+1 \geq \frac{n-\lambda\beta+1}{\lambda\beta} + 1 \geq \frac{N}{\lambda\beta},$$

and $H_j(\boldsymbol{t}) := d_P^\beta(\boldsymbol{t} - \boldsymbol{t}_j)$. Lemma 2.4.11 yields $H_j(\boldsymbol{t}_j) = 1$,

$$\begin{aligned}
|H_j(\boldsymbol{t})| &\leq \left(\frac{1}{2(P+1)\,|\boldsymbol{t}-\boldsymbol{t}_j|_{\mathbb{T}^d}}\right)^\beta \leq \left(\frac{\lambda\beta}{2N\,|\boldsymbol{t}-\boldsymbol{t}_j|_{\mathbb{T}^d}}\right)^\beta, && \boldsymbol{t} \neq \boldsymbol{t}_j,\\
\left|\langle H_k, H_j\rangle_{L^2(\mathbb{T}^d)}\right| &\leq \left\|d_P^\beta\right\|_{L^2(\mathbb{T}^d)}^2 \leq \frac{1}{(P+1)^d\beta^{d/2}} \leq \frac{\lambda^d\beta^{d/2}}{N^d}, && \boldsymbol{t}_k \in \Lambda_l,\\
\left|\langle H_k, H_j\rangle_{L^2(\mathbb{T}^d)}\right| &= \left|\left\langle d_P^\beta, d_P^\beta(\cdot - (\boldsymbol{t}_j - \boldsymbol{t}_k)\right\rangle_{L^2(\mathbb{T}^d)}\right| \leq \frac{\lambda^d\beta^{(d+1)/2}(\lambda\beta)^\beta}{2N^d(N\,|\boldsymbol{t}_j-\boldsymbol{t}_k|_{\mathbb{T}^d})^\beta}, && \boldsymbol{t}_k \notin \Lambda_l.
\end{aligned}$$

Finally, we define $I_j(\boldsymbol{t}) := G_j(\boldsymbol{t})H_j(\boldsymbol{t})$. This yields $I_j \in \mathcal{P}(n)$ since $G_j \in \mathcal{P}(Q(\lambda-1))$, $H_j \in \mathcal{P}(P\beta)$, and

$$P\beta + (\lambda-1)Q \leq \frac{n}{\lambda} + (\lambda-1)\frac{n}{\lambda} = n.$$

Moreover, this function has the desired property $I_j(\boldsymbol{t}_k) = \delta_{jk}$ for all $\boldsymbol{t}_k \in \Lambda_l$ and the two remaining inequalities follow by $|I_j(\boldsymbol{t})| \leq \|G_j\|_{\mathcal{C}(\mathbb{T}^d)}\,|H_j(\boldsymbol{t})|$ and by using $\mathrm{e}^{-\pi i\beta P(\boldsymbol{t}-\boldsymbol{t}_j)}H_j(t) \geq 0$, also $\left|\langle I_k, I_j\rangle_{L^2(\mathbb{T}^d)}\right| \leq \|G_j\|_{\mathcal{C}(\mathbb{T}^d)}^2\left|\langle H_k, H_j\rangle_{L^2(\mathbb{T}^d)}\right|$. □

Remark 3.4.11.
Following the calculation in [73, p. 147], we can improve (3.4.8) *to*

$$\|G_j\|_{\mathcal{C}(\mathbb{T}^d)} \leq \left(1 - \frac{\pi^2}{3\lambda^2}\right)^{\frac{1-\lambda}{2}} \left(\frac{N}{\lambda}/\left\lfloor\frac{n}{\lambda}\right\rfloor\right)^{\lambda-1} \left(\frac{\lambda}{\pi}\right)^{\lambda-1}\mathcal{C} \leq 2.4\left(\frac{C(n)\lambda}{\pi}\right)^{\lambda-1}\mathcal{C}$$

with $C(n) \to 1$ for $n \to \infty$ and where the first two bracketed terms are due to (3.4.3) *and* (3.4.7)*, respectively.*

With the Lagrange-like basis at hand we can now utilize the robust duality and prove a lower bound for the smallest singular value of multivariate Vandermonde matrices.

Theorem 3.4.12 (Lower bound on the smallest singular value)**.**
Let $\beta, d, N, M \in \mathbb{N}$, $\beta \geq d+1$ even, $\Omega = \{\boldsymbol{t}_1, \dots, \boldsymbol{t}_M\} \subset [0,1)^d$ be a clustered node configuration and $N > 2\beta^2\lambda$. Moreover, assume the cluster separation

$$\Delta N \geq \lambda\beta \left(\beta^{1/2} 2^d (2^d - 1)\lambda^\lambda \zeta(\beta - d + 1)\mathcal{C}\right)^{\frac{1}{\beta}}, \tag{3.4.9}$$

where ζ denotes the Riemann zeta function. Then the smallest singular value of the Vandermonde matrix $\boldsymbol{A} \in \mathbb{C}^{M \times N^d}$ from (3.4.1) is bounded by

$$\sigma_{\min}(\boldsymbol{A}) \geq \left(1.5 \cdot \beta^{d/4} \lambda^{\lambda + d/2 - 1/2}\right)^{-1} \frac{N^{d/2}}{\mathcal{C}}.$$

Proof. We apply the robust duality from Lemma 3.4.8, with $\boldsymbol{v} \in \mathbb{C}^M$, $\|\boldsymbol{v}\|_2 = 1$, such that $\sigma_{\min}(\boldsymbol{A}) = \|\boldsymbol{A}^* \boldsymbol{v}\|_2$, and

$$f := \sum_{k=1}^{M} v_k I_k,$$

where the Lagrange-like basis functions I_k are given by Lemma 3.4.10. The interpolation errors $\epsilon_j = f(\boldsymbol{t}_j) - v_j$ fulfill $\boldsymbol{\epsilon} = \boldsymbol{B}\boldsymbol{v}$, where $\boldsymbol{B} \in \mathbb{C}^{M \times M}$ has the entries

$$B_{j,k} := \begin{cases} 0, & j = k, \\ I_k(\boldsymbol{t}_j), & j \neq k. \end{cases}$$

Let $l(k)$ denote the index of the cluster to which the node $\boldsymbol{t}_k$ belongs to. We proceed by $\|\boldsymbol{\epsilon}\|_2 \leq \|\boldsymbol{B}\|_2 \leq \left\|\widetilde{\boldsymbol{B}}\right\|_2$, where the second inequality follows from monotonicity of the norm in Lemma 2.1.23 and Lemma 3.4.10 i) and ii) with

$$\widetilde{B}_{j,k} := \begin{cases} 0, & \boldsymbol{t}_j \in \Lambda_{l(k)}, \\ \frac{\beta^\beta \lambda^{\beta+\lambda-1}}{(2N|\boldsymbol{t}_k - \boldsymbol{t}_j|_{\mathbb{T}^d})^\beta}\mathcal{C}, & \text{otherwise.} \end{cases}$$

Since $\widetilde{\boldsymbol{B}} \in \mathbb{R}^{M \times M}$ is symmetric, we bound the spectral norm by the maximum norm and apply the packing argument from Lemma 3.4.2 to get

$$\begin{aligned} \|\boldsymbol{\epsilon}\|_2 &\leq \max_{j=1,\dots,M} \sum_{\substack{k=1 \\ k \neq j}}^{M} \widetilde{B}_{j,k} \leq \lambda 2^d (2^d - 1) \sum_{m=1}^{\lfloor N/2\rho \rfloor} m^{d-1} \max_{\boldsymbol{t} \in J_m} \frac{\beta^\beta \lambda^{\beta+\lambda-1}}{(2N\,|\boldsymbol{t}|_{\mathbb{T}^d})^\beta}\mathcal{C} \\ &\leq 2^{d-\beta}(2^d - 1)\lambda^{\beta+\lambda}\beta^\beta \mathcal{C}\zeta(\beta - d + 1)(\Delta N)^{-\beta}. \end{aligned}$$

Condition (3.4.9) and $\beta \geq 2$ imply $\|\boldsymbol{\epsilon}\|_2 \leq 1/\sqrt{32}$. To bound the L^2-norm of f, let us define $\hat{\boldsymbol{B}} := \left(|\langle I_k, I_j \rangle|\right)_{j,k=1,\dots,M} \in \mathbb{R}^{M \times M}$. The triangle inequality, symmetry of $\hat{\boldsymbol{B}}$, Lemma 3.4.10 iii), and the packing argument from Lemma 3.4.2 yield

$$\|f\|^2_{L^2(\mathbb{T}^d)} = \sum_{j=1}^{M}\sum_{k=1}^{M} v_k \overline{v}_j \langle I_k, I_j \rangle_{L^2(\mathbb{T}^d)} \leq \max_{\|\boldsymbol{w}\|_2 = 1} \boldsymbol{w}^* \hat{\boldsymbol{B}} \boldsymbol{w} \leq \left\|\hat{\boldsymbol{B}}\right\|_\infty \leq \max_j \sum_{k=1}^{M} \left|\langle I_k, I_j \rangle_{L^2(\mathbb{T}^d)}\right|$$

$$\leq \frac{\lambda^d \beta^{d/2}}{N^d} \lambda^{2\lambda-2} \mathcal{C}^2 \left(\lambda + \lambda 2^d (2^d - 1) \sum_{m=1}^{\lfloor N/2\rho \rfloor} m^{d-1} \max_{\boldsymbol{t} \in J_m} \frac{\sqrt{\beta}}{2} \frac{(\lambda\beta)^\beta}{(N |\boldsymbol{t}|_{\mathbb{T}^d})^\beta} \right)$$

$$\leq \frac{\lambda^d \beta^{d/2}}{N^d} \lambda^{2\lambda-1} \mathcal{C}^2 \left(1 + \lambda^\beta \beta^{\beta+1/2} 2^{d-1} (2^d - 1) \zeta(\beta - d + 1) (\Delta N)^{-\beta} \right).$$

Condition (3.4.9) implies

$$\|f\|_{L^2(\mathbb{T}^d)} \leq \sqrt{\frac{3}{2}} \left(\frac{\lambda\sqrt{\beta}}{N} \right)^{d/2} \lambda^{\lambda - 1/2} \mathcal{C}$$

and with Lemma 3.4.8, we can finish the proof. □

Since the preceding theorem has a lot parameter, we state some corollaries for specific choices of the parameter and situation. For $d = 1$, Lemma 3.4.5 applied to the cluster complexity yields the following.

Corollary 3.4.13 (Bound in terms of minimal separation distance)**.**
Under the assumptions of Theorem 3.4.12 with $d = 1$, $\beta = 2$ and

$$\Delta N \geq 4.4\lambda \left(\frac{2\mathrm{e}}{qN} \right)^{\frac{\lambda-1}{2}},$$

we have

$$\sigma_{\min}(\boldsymbol{A}) \geq \frac{1}{1.8(2\mathrm{e})^{\lambda-1}} \cdot \sqrt{N} (qN)^{\lambda-1}.$$

In Theorem 3.4.12 the parameter β is free and can therefore be used to optimize the result for different purposes. The larger β is chosen, the weaker becomes the bound on the smallest singular value. It also turns out that it can be chosen such that the assumption on the cluster separation depends only logarithmically on the cluster complexity.

Corollary 3.4.14 (Specific choices of β)**.**
Specific choices of β in Theorem 3.4.12 yield the following:

i) By choosing $\beta = d + 1$ or $\beta = d + 2$ for d being odd or even, respectively, and some additional cosmetics, the condition

$$\Delta N \geq 6d\lambda \left(\lambda^\lambda \mathcal{C} \right)^{\frac{1}{d+1}}$$

implies our best estimate

$$\sigma_{\min}(\boldsymbol{A}) \geq \left(3d^{d/4} \lambda^{\lambda + d/2 - 1/2} \right)^{-1} \frac{N^{d/2}}{\mathcal{C}}.$$

ii) By choosing $\beta = 2 \left\lceil \frac{1}{2} \log\left(2^d (2^d - 1) \lambda^\lambda \zeta(2) \mathcal{C} \right) \right\rceil$ and noting that $\sqrt[2\beta]{\beta} \leq 1.2$ for β even and $\sqrt[\log(C)]{C} = \mathrm{e}$, our weakest condition

$$\Delta N \geq 3.3\lambda \left(2.5 + 1.4d + \lambda \log(\lambda) + \log(\mathcal{C}) \right),$$

implies

$$\sigma_{\min}(\boldsymbol{A}) \geq \left(1.5 \cdot (2.5 + 1.4d + \lambda \log(\lambda) + \log(\mathcal{C}))^{d/4} \lambda^{\lambda + d/2 - 1/2} \right)^{-1} \frac{N^{d/2}}{\mathcal{C}}.$$

Corollary 3.4.15 (Pair clusters)**.**
For $\lambda = 2$, we have $\mathcal{C} = \frac{1}{qN}$ and at most pairs of nodes form clusters. Corollary 3.4.14 i) with

$$\Delta N \geq 12d \left(\frac{4}{qN}\right)^{\frac{1}{d+1}}$$

implies

$$\sigma_{\min}(\boldsymbol{A}) \geq \frac{(qN)N^{d/2}}{12 \cdot 2^{d/2-1/2} \cdot d^{d/4}}.$$

Finally, we end this section with a remark concerning the extension to a possibly larger class of nodes.

Remark 3.4.16 (Extended clusters)**.**
Definition 3.4.3 assumes clusters to be a set of nodes contained in an interval of length $1/N$. This constraint can be weakened by means of Lemma 3.5.9. For example, we can allow clusters to be nodes in an interval of length λ/N (also suggested by the experiment in Figure 3.2.1). Albeit, further restrictions to the cluster separation are imposed and a worse constant in the bound on the singular value is obtained. While the statement for usual clustered node configurations is of the form: if $N > a$ and $\Delta N > b$, then

$$\sigma_{\min}(\boldsymbol{A}) \geq N^{d/2} \frac{c}{(qN)^{\lambda-1}}$$

for some constants a, b, c, we have after applying Lemma 3.5.9 the following. Let $N \in \mathbb{N}_+$ and Ω be a clustered node configuration with clusters defined as node sets in an interval of length λ/N. If $N = n\lambda$ for some $n \in \mathbb{N}$, $N > \lambda a$ and $\Delta N > \lambda b$, then

$$\sigma_{\min}(\boldsymbol{A}) \geq N^{d/2} \frac{c}{(\lambda qN)^{\lambda-1}}.$$

This is a factor $\lambda^{d/2}$ better than simply applying the monotony of the singular values from Lemma 2.1.18.

3.4.4 Upper bounds on the smallest singular value and beyond distances

Next, we show that the obtained lower bounds are sharp for $d = 1$ and for $\lambda = 2$, respectively. Moreover, we show for $d > 1$ and nodes in generic position (e.g. not all nodes on a line for $d = 2$), that the cluster complexity C is not the optimal quantity to understand the situation here. If we assume a minimal separation q between nodes, then the estimate in Theorem 3.4.12 is sub-optimal with respect to the order in qN we can derive from the cluster complexity. For this, we give an example with one cluster of three nodes in the bivariate case, $d = 2$. All numerical examples are processed with fixed column parameter N that is specified at the appropriate places.

Example 3.4.17 (Matching bounds for $d = 1$)**.**
In [73, Prop. 2.10] an upper bound on $\sigma_{\min}(\boldsymbol{A})$ is given for a clustered node configuration that consists of at least one cluster of λ equispaced, q-separated nodes. After further simplifications, we can derive

$$\min_{\Omega} \sigma_{\min}(\boldsymbol{A}) \leq (\pi\lambda)^{1/4} \pi^{\lambda-1} \sqrt{N} (qN)^{\lambda-1} (1 + qNC(\lambda)\sqrt{N}), \quad C(\lambda) := 2\pi \sum_{l=0}^{\lambda-1} \binom{\lambda-1}{l} \frac{l^\lambda}{\lambda!}.$$

Together with Remark 3.4.11 and Corollary 3.4.13 this assures that for sufficiently large $N \in \mathbb{N}$, small q and $\lambda \geq 2$, there exist constants $c_1 \leq c_2$ such that

$$\sqrt{N}\,(c_1 qN)^{\lambda-1} \leq \min_{\Omega} \sigma_{\min}(\boldsymbol{A}) \leq \sqrt{N}\,(c_2 qN)^{\lambda-1},$$

where the minimum is taken over all clustered node configurations Ω with at least one cluster of λ nodes with minimal separation q. This was also conjectured in [11, Rem. 3.5]. Moreover, we note that the lower bound in Lemma 3.4.5 implies that the term λ^λ in Theorem 3.4.12 cannot be avoided.

Example 3.4.18 (Matching bounds for $\lambda = 2$)**.**
Let $d \in \mathbb{N}$, $\lambda \geq 2$, and the node set Ω with minimal separation distance $q \leq 1/N$ such that $|\boldsymbol{t}_1 - \boldsymbol{t}_2|_{\mathbb{T}^d} = q$. Then Lemma 2.4.7, the interlacing theorem for eigenvalues (Lemma 2.1.6) and the binomial formula yield

$$\begin{aligned}\sigma_{\min}(\boldsymbol{A})^2 &\leq N^d\,(1 - |d_n\,(\boldsymbol{t}_1 - \boldsymbol{t}_2)|) \leq N^d \left(1 - \left(1 - \frac{\pi^2}{6}(qN)^2\right)^d\right) \\ &\leq N^d \frac{\pi^2 (qN)^2}{6} \sum_{k=0}^{d-1} \left(1 - \frac{\pi^2 (qN)^2}{6}\right)^k \leq \frac{\pi^2 (qN)^2 d N^d}{6}.\end{aligned} \tag{3.4.10}$$

Together with Corollary 3.4.15, there exists constants $c_1(d) \leq c_2(d)$ such that

$$c_1(d) N^{d/2} qN \leq \min_{\Omega} \sigma_{\min}(\boldsymbol{A}) \leq c_2(d) N^{d/2} qN,$$

where the minimum is taken over all clustered node configurations Ω with at least one cluster of $\lambda = 2$ nodes with minimal separation q.

We present a numerical experiment in order to confirm our results for the higher dimensional case and set $d = 2$. Randomized clustered node configurations of $S = 2$, $S = 20$ and $S = 40$ clusters with 2 nodes each are constructed for 100 different minimal separations q, respectively. Then the smallest singular values of the corresponding Vandermonde matrices $\sigma_{\min}(\boldsymbol{A})$ are computed and the upper bound from (3.4.10) and the lower bound from Corollary 3.4.15 are shown. The results are presented in Figure 3.4.3. The node configurations are built as follows. The minimal separation q is picked logarithmically uniformly at random in $[10^{-3}/N, 1/N]$. We set $N = 10^4$ so that the condition on Δ in Corollary 3.4.15 together with the left interval bound for q make $\Delta \geq \Delta_{\min}$ (value shown in the figure) necessary. Two clusters realize the cluster separation $\Delta_{\min}$ and for the remaining clusters, we pick a position in $[0,1]^2$ uniformly at random. The positions are fixed for the respective choice of S and do not change for different q. Each cluster is constructed randomly by setting one node to $(0,0)$ and one to either $(a,1)$ or $(1,a)$ for some $a \in [0,1]$. Then we scale the clusters by q and move them to their respective cluster positions.

Example 3.4.19 (Triple cluster)**.**
Let $d = 2$, $N \in \mathbb{N}$, and $\Omega = \{\boldsymbol{t}_1, \boldsymbol{t}_2, \boldsymbol{t}_3\} \subset [0,1)^2$ with

$$\boldsymbol{t}_1 = \begin{pmatrix}0\\0\end{pmatrix},\ \boldsymbol{t}_2 = \eta\begin{pmatrix}a_1\\a_2\end{pmatrix},\ \boldsymbol{t}_3 = \eta\begin{pmatrix}b_1\\b_2\end{pmatrix},\quad \eta \in \left(0, \frac{1}{2}\right],\ a_1^2 + a_2^2 = b_1^2 + b_2^2 = 1,\ a_1 b_1 + a_2 b_2 \leq 0.$$

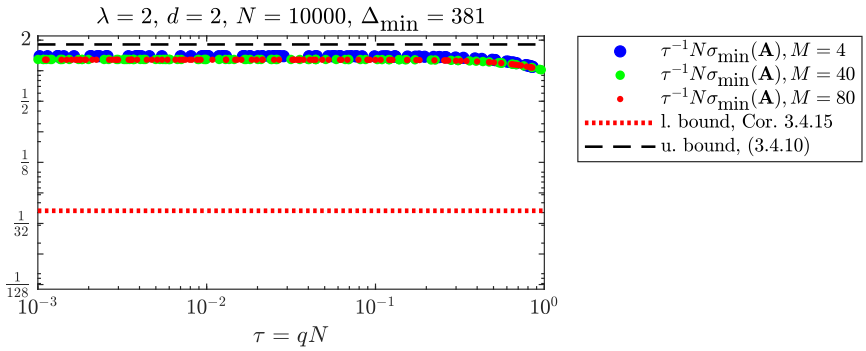


Figure 3.4.3: upper and lower bounds on $\sigma_{\min}(\boldsymbol{A})$ for bivariate pair clusters as in Corollary 3.4.15 and Example 3.4.18.

and hence, the minimal separation of Ω *is* $\eta/\sqrt{2} \leq q \leq \eta$. *Then the smallest singular value of the corresponding Vandermonde matrix* $\boldsymbol{A}$ *fulfills*

$$\sigma_{\min}(\boldsymbol{A}) = \begin{cases} \Theta((qN)^2), & \text{antipodal nodes, } a_1b_1 + a_2b_2 = -1, \\ \Theta(qN), & \text{otherwise,} \end{cases}$$

and this can be seen as follows: Define the real matrix

$$\boldsymbol{M} := \begin{pmatrix} 1 & u & v \\ u & 1 & w \\ v & w & 1 \end{pmatrix} := \left(\mathrm{e}^{-\pi i m(\boldsymbol{t}_j - \boldsymbol{t}_k)} d_m(\boldsymbol{t}_j - \boldsymbol{t}_k) \right)_{j,k=1,2,3},$$

note that $\sigma_{\min}(\boldsymbol{A})^2 = \sigma_{\min}(\boldsymbol{A}\boldsymbol{A}^*) = \sigma_{\min}(\boldsymbol{M}) = \left\| \boldsymbol{M}^{-1} \right\|_2^{-1}$, *and use the explicit formula*

$$\left\| \boldsymbol{M}^{-1} \right\|_2 = \frac{1}{|u^2 + v^2 + w^2 - 2uvw - 1|} \left\| \begin{pmatrix} w^2 - 1 & u - vw & v - uw \\ u - vw & v^2 - 1 & w - uv \\ v - uw & w - uv & u^2 - 1 \end{pmatrix} \right\|_2 . \tag{3.4.11}$$

The univariate Taylor expansion

$$\mathrm{e}^{-\pi i n \eta} d_n(\eta) = 1 - \alpha_n (\eta N)^2 + \gamma_n (\eta N)^4 + \mathcal{O}\left((\eta N)^6\right), \quad \alpha_n, \gamma_n \neq 0,$$

and $a_1^2 + a_2^2 = 1 = b_1^2 + b_2^2$ *yield*

$$u = 1 - \alpha_n (\eta N)^2 + (\alpha_n^2 a_1^2 a_2^2 + \gamma_n (a_1^4 + a_2^4))(\eta N)^4 + \mathcal{O}\left((\eta N)^6\right)$$

and similar expressions for the other quantities. By direct computation, we see that the entries in the matrix on the right hand side of (3.4.11) *are all* $\mathcal{O}\left((\eta N)^2\right)$ *and for example the diagonal entry* $u^2 - 1$ *is* $\Theta((\eta N)^2)$ *independent of* $\boldsymbol{a}$ *and* $\boldsymbol{b}$. *Hence, the norm of that matrix is* $\Theta((\eta N)^2)$. *Similarly, the denominator of* (3.4.11) *can be computed to be*

$$u^2 + v^2 + w^2 - 2uvw - 1 = \begin{cases} \mathcal{O}\left((\eta N)^6\right), & a_1b_1 + a_2b_2 = -1, \\ \Theta((\eta N)^4), & \text{otherwise.} \end{cases}$$

Finally, this yields

$$\sigma_{\min}(\boldsymbol{A}) = \begin{cases} \mathcal{O}\left((\eta N)^2\right), & a_1b_1 + a_2b_2 = -1, \\ \Theta((\eta N)), & \text{otherwise,} \end{cases}$$

and together with Theorem 3.4.12 the assertion.

We present a numerical experiment for this phenomenon. We set $N = 100$, $d = 2$ *and build the triple cluster consisting of the nodes* $\boldsymbol{t}_1 = (0,0)^\top$, $\boldsymbol{t}_2 = (-\sqrt{1-a^2}\eta, a\eta)^\top$ *and* $\boldsymbol{t}_3 = (\eta, 0)^\top$ *(see Figure 3.4.4, left), where* $q = \eta\sqrt{1-a^2} \in [10^{-6}, 1/2]$ *is picked logarithmically uniformly at random. Then we compute the smallest singular value of the Vandermonde matrix* $\sigma_{\min}(\boldsymbol{A})$. *This is repeated* 100 *times for* $a = 0.01$ *and* $a = 0$ *each. The results are presented in Figure 3.4.4 (right). We see the asymptotic behavior with respect to* qN *calculated in Example 3.4.19. Furthermore, for nodes not being antipodal, we observe that the asymptotic starts when* qN *becomes smaller than the displacement* a.

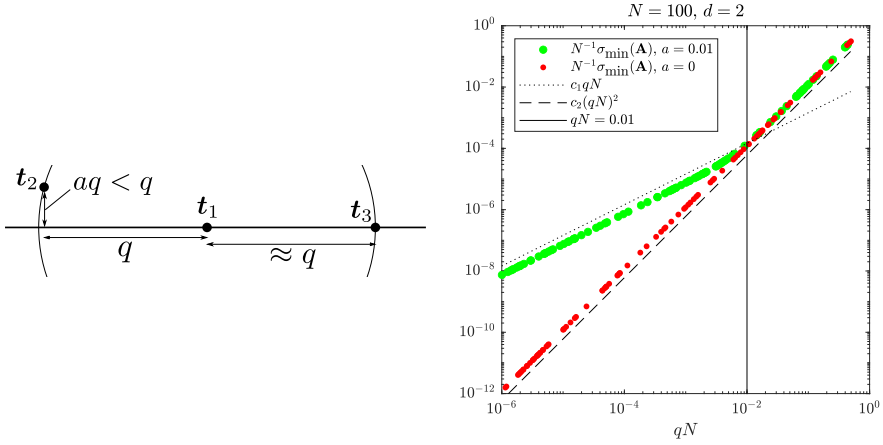


Figure 3.4.4: triple cluster, almost antipodal nodes, cf. Example 3.4.19; left: sketch of node positions; right: numerical results.

3.4.5 Comparison to univariate results for clustered nodes

In this section we state results from the literature that are closely related to ours, in particular the results from [11, 73, 34, 12] that came up during the research time for this thesis. They all deal with clustered node configurations in one dimension only. We make appropriate comparisons between these and Theorem 3.4.12 for $d = 1$ in different examples. All numerical examples are processed with fixed column parameter N that is specified at the appropriate places. First of all we compare our theorem to [73] that provides the proof technique we used.

Theorem 3.4.20 ([73, Thm. 2.3])**.**
Let $d = 1$ $M, N \in \mathbb{N}_+, M < N$ *and suppose* $\Omega \subset \mathbb{T}$ *is a clustered node configuration with* S *clusters* $\{\Lambda_\ell\}_{\ell=1}^S$ *of cardinality* $\lambda_1 \geq \cdots \geq \lambda_S$ *(contained in an interval of length* $\frac{1}{N-1}$*). We define the node specific complexity by*

$$\mathcal{C}_j := \mathcal{C}(t_j) := \prod_{t_j \in \Omega:\ 0 < |t_j - t'|_{\mathbb{T}^d} \leq 1/(N-1)} \frac{1}{\pi(N-1)\,|t_j - t'|_{\mathbb{T}}}.$$

Assume that $N \geq 2M^2$ *and*

$$\Delta(N-1) \geq \max_{1\leq \ell \leq S} \max_{t_j \in \Lambda_\ell} 10\lambda_\ell^{5/2} (M\mathcal{C}_j)^{\frac{1}{2\lambda_\ell}}.$$

Then we have for the univariate Vandermonde matrix $\boldsymbol{A} \in \mathbb{C}^{M\times N}$

$$\sigma_{\min}(\boldsymbol{A}) \geq \sqrt{N-1}\left(\sum_{\ell=1}^{S}\sum_{t_j\in\Lambda_\ell}\left(b_\ell \lambda_\ell^{\lambda_\ell}\mathcal{C}_\ell\right)^2\right)^{-\frac{1}{2}} {}^{*},$$

where the constant b_ℓ *is given by*

$$b_\ell := \frac{20\sqrt{2}}{19}\left(1-\frac{\pi^2}{3\lambda_\ell^2}\right)^{-\frac{\lambda_\ell-1}{2}}\left(\frac{N-1}{\lambda_\ell}\right)^{\lambda_\ell-1}\left\lfloor\frac{N-1}{\lambda_\ell}\right\rfloor^{-\lambda_\ell-1}.$$

Example 3.4.21 (Comparison with [73])**.**
Let $d=1$ *and* $\beta=2\lambda$*, then* $N>2\lambda^3$ *and* $\Delta N \geq 4.4\lambda^{5/2}\mathcal{C}^{\frac{1}{2\lambda}}$ *imply with Theorem 3.4.12*

$$\sigma_{\min}(\boldsymbol{A}) \geq \left(1.8\cdot C_0^{\lambda-1}\cdot\lambda^{\lambda+1/4}\right)^{-1}\frac{\sqrt{N}}{\mathcal{C}}, \tag{3.4.12}$$

where we set $C_0=1$ *for the moment. This can be compared to Theorem 3.4.20, where up to re-normalization to* N*,* $N>2\lambda^2$ *and* $\Delta N \geq 10\lambda^{5/2}(M\mathcal{C})^{\frac{1}{2\lambda}}$ *imply*

$$\sigma_{\min}(\boldsymbol{A}) \geq \left(1.5\cdot C_0^{\lambda-1}\cdot\sqrt{M}\lambda^{\lambda}\right)^{-1}\frac{\sqrt{N}}{\mathcal{C}}. \tag{3.4.13}$$

According to Remark 3.4.11, $C_0\in\left(\pi^{-1},1\right]$ *depending on* λ *and* $N-1$*. In total, we have a stronger condition on* N*, but our condition on* Δ *is always weaker and our estimate on* $\sigma_{\min}(\boldsymbol{A})$ *is sharper if* $M>2$*.*

The following numerical example serves to compare Theorem 3.4.12 with Theorem 3.4.20 in the univariate case, $d=1$*, for bigger clusters. The column parameter is set to* $N=2^{15}$*. We build up clustered node configurations with* $S=2$ $(M=10)$ *and* $S=10$ $(M=50)$ *clusters of size* $\lambda=5$ *placed equispaced at* $\frac{l}{S}$ *for* $l=0,\dots,S-1$ *At each cluster position the cluster nodes start to lie equispaced with separation* q*, where* $q\in[10^{-4}/N, 1/(4N)]$ *(the right hand interval bound is due to a cluster is lying in an interval of length* $1/N$*) is picked logarithmically uniformly at random. Afterwards the smallest singular value* $\sigma_{\min}(\boldsymbol{A})$ *is computed. This procedure is repeated 100 times for the respective choice of* S *and the results are presented in Figure 3.4.5. We use the statements from Example 3.4.21 with* $C_0=(1-\frac{\pi^2}{3\lambda^2})^{-1/2}N/\lambda\lfloor n/\lambda\rfloor^{-1}$*. Since* $d=1$*, the worst case cluster complexity is estimated by* (3.4.4) *to* $\mathcal{C}\leq(qN)^{-4}/4$*.*

The next theorem is from [11] and it is the first result proving that the smallest singular value of $\boldsymbol{A}$ corresponding to clustered node configurations has a lower bound with scaling $(qN)^{\lambda-1}$. This result replaced the former best bound with scaling $(qN)^{M-1}$. However, it is derived from a continuous result by discretization and comes therefore with huge technical restrictions and quite pessimistic constants.

Theorem 3.4.22 (cf. [11, Cor. 3.6])**.**
Let $N=2n+1\in\mathbb{N}_+$*,* $d=1$ *and* $\Omega\subset\mathbb{T}$ *contained in an interval of length* $\frac{1}{2M^2}$ *be a clustered node configuration with minimal cluster separation satisfying*

$$\Delta(N-1)\geq 4M.$$

*In the referred manuscript the constant $\lambda_\ell^{\lambda_\ell-1}$ is given, which is not correct. On page 148 in [73] a factor λ_a (with the notation in there) is missing on the right hand side of the third formula.

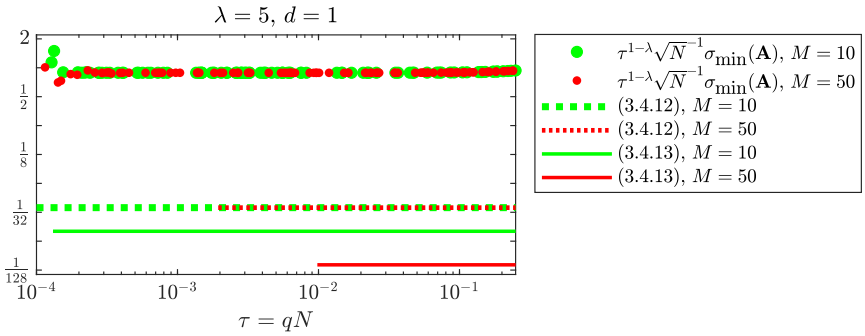


Figure 3.4.5: node configurations with bigger clusters. Lower bounds on $\sigma_{\min}(\boldsymbol{A})$. Comparison as presented in Example 3.4.21 with estimate from Remark 3.4.11.

Then the smallest singular value of the accompanying univariate Vandermonde matrix $\boldsymbol{A} \in \mathbb{C}^{M\times N}$ *fulfills*

$$\sigma_{\min}(\boldsymbol{A}) \geq \frac{\sqrt{N-1}}{2(2\pi)^{M-1}M^{2M-1}} \left(\pi q(N-1)\right)^{\lambda-1}.$$

This result holds also with redefined clusters that are nodes contained in an interval of length λ/N.

We compare this result to Theorem 3.4.12 in the case when all nodes are one cluster because this situation is where the assumptions of Theorem 3.4.22 are least restrictive.

Example 3.4.23 (All nodes cluster, comparison to Theorem 3.4.22)**.**
Let $d = 1$ *and* Ω *be a cluster, i.e.* $\lambda = M$*. If* $N > 8M$*, then Corollary 3.4.13 implies*

$$\sigma_{\min}(\boldsymbol{A}) \geq \frac{1}{1.8(2\mathrm{e})^{M-1}} \cdot \sqrt{N}(qN)^{M-1}.$$

This compares to Theorem 3.4.22, where restricting the nodes to an interval of length $1/(2M^2)$ *and* $N \geq 4M^3$ *imply*

$$\sigma_{\min}(\boldsymbol{A}) \geq \frac{1}{2^M M^{2M-1}} \cdot \sqrt{N}(qN)^{M-1},$$

but, note that the definition of a clustered node configuration in [11] is in principle more flexible than ours, since the clusters can be more extended.

The following two theorems are from [12] and describe the behavior of the full spectrum of a Vandermonde matrix with clustered nodes. They become effective in the regime of clusters being in an interval a lot smaller than $1/N$. The first theorem shows that each singular value of $\boldsymbol{A}$ is comparable to one singular value of a submatrix of $\boldsymbol{A}$ that corresponds to the nodes of one cluster (taking only the rows of $\boldsymbol{A}$ belonging to the cluster). The second theorem provides a description of the whole spectrum of a Vandermonde matrix corresponding to a node subset that builds one cluster. Compared to our results, the focus lies on the cluster width rather than the smallest separation distance.

Theorem 3.4.24 (cf. [12, Thm. 2.2])**.**
Let $M, N \in \mathbb{N}_+, N \geq M$*,* $d = 1$ *and* $\Omega \subset \mathbb{T}$ *be a node configuration with clusters* $\{\Lambda_\ell\}_{\ell=1}^{S}$ *of* λ_ℓ *nodes, cluster separation* $\Delta > 1/N$ *and extent* $h_\ell := \max_{t,t'\in\Lambda_\ell} |t-t'|_{\mathbb{T}}$ *at most* $h \leq 1/N$*.*

Let $\boldsymbol{A}(\Lambda_\ell) \in \mathbb{C}^{\lambda_\ell \times N}$ denote the Vandermonde matrix corresponding to the nodes in the cluster Λ_ℓ and

$$\tilde{\sigma}_1 \geq \cdots \geq \tilde{\sigma}_M$$

the ordered singular values of all $\boldsymbol{A}(\Lambda_\ell), \ell = 1, \ldots, S$. Then there exists positive constants C_1, C_2, C_3, C_4 only depending on λ_j, such that for all $\frac{C_1}{\Delta} \leq N \leq \frac{C_2}{h}$ the singular values of the full matrix $\boldsymbol{A}$ fulfill

$$\left(1 - \frac{C_3}{N\Delta} - C_4 Nh\right)^{\frac{1}{2}} \tilde{\sigma}_j \leq \sigma_j(\boldsymbol{A}) \leq \tilde{\sigma}_j \left(1 + \frac{C_3}{N\Delta} + C_4 Nh\right)^{\frac{1}{2}}, \qquad j = 1, \ldots, M.$$

Theorem 3.4.25 (cf. [12, Thm. 2.3])**.**
Let $M, N \in \mathbb{N}_+$, $d = 1$ and Ω be a cluster of M nodes with minimal separation q and extent h. Then there exist constants $C_5(q, M)$, $C_6(q, M)$ and $C_7(M)$, such that for all $N > M$ and $Nh \leq C_5$ we have

$$C_6 N^{1/2} (hN)^{M-j} \leq \sigma_j(\boldsymbol{A}) \leq C_7 N^{1/2} (hN)^{M-j}, \qquad j = 1, \ldots, M.$$

Combining both theorems and assuming that the nodes inside the clusters are $q = h/c$ separated for some constant $c \geq \lambda_j$ (almost equispaced) the scaling of each singular value of $\boldsymbol{A}$ with respect to qN can be described. We can conclude (provided the technical assumptions are fulfilled) that for each cluster Λ_j there exists exactly one singular value of $\boldsymbol{A}$ that scales like $(qN)^{\lambda_j - 1}$, where λ_j is the size of that cluster. In particular this affirms the notion from previous sections that the smallest singular value is caused by the largest cluster. Since the cluster extent h is in the case of pair clusters qc with uniformity parameter c in (3.3.9), it offers a review and a comparison to this technique in that setting.

Remark 3.4.26 (Comparison to Theorem 3.4.24 for pair clusters)**.**
To prove Theorem 3.4.24 Batenkov et al. used a QR-decomposition technique. Adapted to the case of pair clusters, we obtain the following, the details are given in Appendix B: Let $M \geq 4$ even and $\boldsymbol{A}$ as in (3.1.6). With respect to the pair clusters, partition $\boldsymbol{A}^* = (\boldsymbol{A}_1^* \ \boldsymbol{A}_2^* \ \ldots \ \boldsymbol{A}_{M/2}^*)$ with QR-decompositions $\boldsymbol{A}_j^* = \boldsymbol{Q}_j \boldsymbol{R}_j$ and set $\boldsymbol{Q} := (\boldsymbol{Q}_1 \ \boldsymbol{Q}_2 \ \ldots \ \boldsymbol{Q}_{M/2})$. Tracing back all constants in lemmata and proofs for the case of pairwise nearly-colliding nodes, we obtain the uniform off-diagonal estimate

$$|(\boldsymbol{Q}^* \boldsymbol{Q})_{j,k}| \leq \frac{150}{\Delta N} + 1079 qN, \qquad j \neq k,$$

which yields a constant "multiplicative perturbation" in [12, Lem. 5.1] and thus a condition number estimate like Theorem 3.3.13 only if $qN \leq C_1/M$ and $C_2 M \leq \Delta N$, for some constants C_1, C_2.

However, note that for two pair clusters $u_1 < u_2 \ll v_1 < v_2$, a direct computation (avoiding a so-called limit basis used in [12]) yields the off-diagonal estimate

$$\|\boldsymbol{Q}_1^* \boldsymbol{Q}_2\|_{\mathrm{F}} \leq \frac{116}{N(v_1 - u_2)}, \qquad \boldsymbol{Q}_1 = \boldsymbol{Q}_1(u_1, u_2), \ \boldsymbol{Q}_2 = \boldsymbol{Q}_2(v_1, v_2).$$

Together with $\Delta N \geq \frac{27}{23} \cdot 232(\log \lfloor \frac{M}{4} \rfloor + 1)$ and Lemma 2.1.22 this gives

$$|1 - \lambda_r(\boldsymbol{Q}^* \boldsymbol{Q})| \leq \max_j \sum_{\ell=1}^{M/2} \|\boldsymbol{Q}_j^* \boldsymbol{Q}_\ell\|_{\mathrm{F}} \leq 2 \sum_{\ell=1}^{\lfloor M/4 \rfloor} \frac{116}{\ell \rho} \leq \frac{232(\log \lfloor \frac{M}{4} \rfloor + 1)}{\rho} \leq \frac{23}{27}, \quad r = 1, \ldots, M.$$

Lemma 2.1.16 finally yields

$$\operatorname{cond}(\boldsymbol{A}) \leq \operatorname{cond}(\boldsymbol{Q}) \cdot \max_j \operatorname{cond}(\boldsymbol{A}_j) \leq \frac{5}{qN}.$$

Altogether, the improved variant of this technique can be used for pair clusters, but leads to a stronger assumption on Δ for all moderate uniformity constants c compared to Theorem 3.3.13.

The following result generalizes the technique we used to proof Theorem 3.1.7 for the case of pair clusters. This is the most superior approach for this situation.

Theorem 3.4.27 (Lower bound for pair clusters, [34])**.**
Let $d = 1$. If $\Omega \subset \mathbb{T}$ consists of clusters that are contained in an interval of length smaller than $3/N$ with $\lambda \leq 2$, cluster separation $\Delta \geq 3/N$ and minimal separation distance q, then we have for the associated Vandermonde matrix $\boldsymbol{A}$

$$\sigma_{\min}(\boldsymbol{A}) \geq \frac{\pi^2}{2 \cdot 3^5}(N+1)^3 q^3.$$

Remark 3.4.28 (Best lower bound for pair clusters)**.**
In Theorem 3.4.27 slightly broader clusters are allowed than we demand by our definition of a clustered node configuration (Definition 3.2.2). Especially, there is no gap between two nodes building a cluster or being well-separated from each other. However, this leads to a weaker constant in the bound of $\sigma_{\min}(\boldsymbol{A})$ applied to our setting and compared to our bound from the proof of Theorem 3.3.13 ($1/\sqrt{11.3}$). Adapting the proof in [34, Thm. 3.6] to our setting (particularly using $qN \leq 1$) and improving estimates in [34, Eq. (8)] provides the best lower bound for pair clusters so far: If $\Omega \subset \mathbb{T}$ is a clustered node configuration with at most pair clusters that are separated by $\Delta N \geq 3$, then

$$\sigma_{\min}(\boldsymbol{A}) \geq \sqrt{N}\frac{\pi}{\sqrt{27}}qN. \tag{3.4.14}$$

We conclude by presenting a comparison between most of the results in the situation of pair clusters.

Example 3.4.29 (Pair clusters, comparison)**.**
Let $d = 1$ and $\lambda = 2$. We apply Theorem 3.4.12 with $\beta = 2$, $\beta = 2\left\lceil \frac{1}{2}\log\left(\frac{\pi^2}{3}\lambda^\lambda \mathcal{C}\right)\right\rceil$ and $\beta = 2\lambda$, respectively. These results are compared in Table 3.4.1 to Theorem 3.4.20 (where we simplified slightly $\lfloor N - 1/\lambda \rfloor \approx N - 1/\lambda$), to Theorem 3.3.13 (under the additional assumption that all nodes inside the clusters have the same separation), and to (3.4.14).

Ref.	Thm. 3.4.12			Thm. 3.4.20	Thm. 3.3.13	(3.4.14)
$\Delta N \geq$	$\frac{17.3}{\sqrt{qN}}$	$34.9 + 6.6\,\lvert\log(qN)\rvert$	$\frac{29}{\sqrt[4]{qN}}$	$\frac{42.5\sqrt[4]{M}}{\sqrt[4]{qN}}$	$25(\log(\lfloor \frac{M}{4} \rfloor) + 1)$	3
$\sigma_{\min}(\boldsymbol{A}) \geq$	$\frac{qN\sqrt{N}}{7.2}$	$\frac{qN\sqrt{N}}{6\sqrt[4]{5.3+\lvert\log(qN)\rvert}}$	$\frac{qN\sqrt{N}}{8.6}$	$\frac{qN\sqrt{N}}{4.5\sqrt{M}}$	$\frac{qN\sqrt{N}}{3.5}$	$\frac{qN\sqrt{N}}{1.7}$

Table 3.4.1: comparison of conditions on the cluster separation Δ and of lower bounds on the singular value $\sigma_{\min}(\boldsymbol{A})$ for the univariate case $d = 1$ with pair clusters, $\lambda = 2$.

We present a numerical experiment for this comparison, set $d = 1$, $N = 2^{15} + 1$ *(Theorem 3.3.13 requires odd* N *without further considerations), and take* $M = 4$ *and* $M = 20$ *nodes, respectively. The node configuration consists of uniformly placed clusters (at* l/N, $l = 1, \dots, M/2$*) that include two nodes each. The first cluster realizes the minimal separation* q*, which is picked logarithmically uniformly at random from* $[10^{-12}/N, 1/N]$*, i.e.* $t_1 = 0$ *and* $t_2 = q$*. The further clusters have nodes* $t_{2l} = l/N$ *and* $t_{2l+1} = l/N + \eta$ *for* $l = 1, \dots, (M-1)/2$*, where* $\eta \in [q, 2q]$ *(parameter* $c = 2$ *in Theorem 3.3.13) is picked uniformly randomly. Afterwards, we compute* $\sigma_{\min}(\boldsymbol{A})$*, where* $\boldsymbol{A}$ *is the Vandermonde matrix defined in* (3.1.6) *corresponding to the node configuration. For each* M *we pick* 50 *instances of* q *and the results are presented in Figure 3.4.6. This clustered node configuration fulfills* $\Delta N \geq \frac{2N}{M} - 1$ *independently of* q*. Theorem 3.4.12 and Theorem 3.4.20 make restrictions to* q *through the condition on* Δ*. Therefore, choosing* β *logarithmically as in Corollary 3.4.14 ii) requires* $qN \geq \mathrm{e}^{-\frac{35.9-2N/M}{6.6}}$*, which is below* 10^{-200} *for both* $M = 4$ *and* $M = 20$*. Theorem 3.4.20 and our result, Theorem 3.4.12, with* $\beta = 4$ *requires respectively*

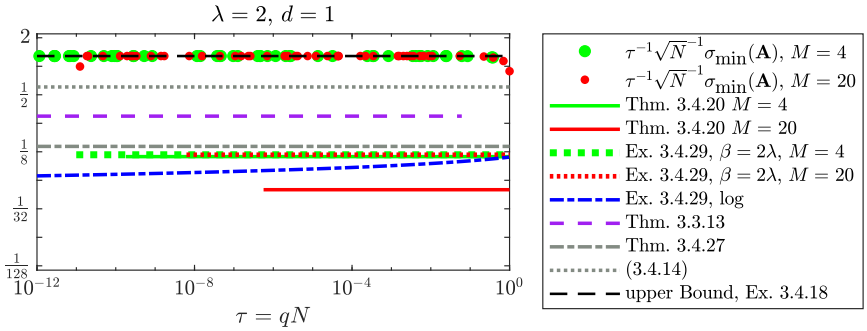


Figure 3.4.6: comparison of different results for the case of pair clusters as in Example 3.4.29.

$$qN \geq \frac{43^4 M}{(\Delta N)^4} \approx \begin{cases} 1.9 \cdot 10^{-10}, & M = 4, \\ 5.9 \cdot 10^{-7}, & M = 20, \end{cases} \quad \text{and} \quad qN \geq \frac{29^4}{(\Delta N)^4} \approx \begin{cases} 9.8 \cdot 10^{-12}, & M = 4, \\ 6.1 \cdot 10^{-9}, & M = 20. \end{cases}$$

3.5 Multivariate well-separated nodes

This last section of the chapter is devoted once again well-separated node sets, but this time multivariate.

Definition 3.5.1 (Well-separated node set, multivariate)**.**
Let $N, M, d \in \mathbb{N}_+$ *and* $\boldsymbol{A} \in \mathbb{C}^{M \times N^d}$ *a Vandermonde matrix with nodes* $\Omega \subset \mathbb{T}^d$*, see* (3.4.1)*. The node set* Ω *is called* well-separated *if the minimal separation distance* q *fulfills*

$$q > \frac{1}{N},$$

see Figure 3.5.1 for an illustration.

Remark 3.5.2.
In order to call a node set well-separated, there only need to be one coordinate well-separated for each distinct pair of nodes. Furthermore, as mentioned in Definition 3.4.3 ii) the class of

well-separated node sets is contained in the class of clustered node configuration, since each well-separated node set is a clustered node configuration with largest cluster size $\lambda = 1$.

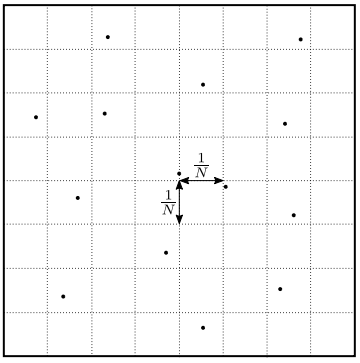


Figure 3.5.1: well-separated bivariate node set in $\mathbb{T}^2$.

There is only few work done for this setting and we give a short summary. The work done in [71] mainly deals with multivariate kernel matrices of the form $\boldsymbol{ACA}^*$ for some appropriate diagonal weight matrix $\boldsymbol{C}$ leading to powers of the Dirichlet kernel evaluated at the node distances as entries. The extremal eigenvalues of such matrices are studied and the lower bound for the smallest eigenvalues can in principle be transferred to lower bounds on $\lambda_{\min}(\boldsymbol{A})$ by Lemma 2.1.17 or Lemma 3.1.6. We present this similarly in Theorem 3.5.13 but with the modified Dirichlet kernel powers from Definition 2.4.10.

If nodes are a full set of equispaced nodes, then the kernel matrix $\boldsymbol{AA}^*$ is a multilevel circulant matrix (see [89, Sec. 3.3] for a definition of a circulant matrix) which gets diagonalized by the discrete, multivariate Fourier matrix. Therefore, by diagonalizing, its eigenvalues become explicit and from that the following theorem is derived.

Theorem 3.5.3 (Equispaced nodes, cf. [71, Cor. 4.11])**.**
For $M \geq 2$*, let* $\Omega = \{\boldsymbol{t}_1, \dots, \boldsymbol{t}_M\} \subset \mathbb{T}^d$ *be a set of equispaced nodes and* $N \in \mathbb{N}_+$ *be even such that* $N > \frac{1}{q} = M^{d/2}$*. Then we have for the Vandermonde matrix* $\boldsymbol{A} \in \mathbb{C}^{M \times N^d}$

$$\left(N - \frac{1}{q}\right)^{\frac{d}{2}} \leq N^{\frac{d}{2}} \left(\frac{\lfloor Nq \rfloor}{Nq}\right)^{\frac{d}{2}} = \sigma_{\min}(\boldsymbol{A}) \leq N^{\frac{d}{2}} \leq \sigma_{\max}(\boldsymbol{A}) = N^{\frac{d}{2}} \left(\frac{\lceil Nq \rceil}{Nq}\right)^{\frac{d}{2}} \leq \left(N + \frac{1}{q}\right)^{\frac{d}{2}}.$$

As a direct consequence we obtain the following corollary for nodes on the grid.

Corollary 3.5.4 (Nodes on the grid)**.**
Let $N \in \mathbb{N}_+$ *be even and* $\Omega \subset \mathbb{T}^d$ *be a set of nodes on the grid with parameter* $\gamma \in \mathbb{N}_+, N > \gamma$*, i.e. the grid width is* $\frac{1}{\gamma}$*. Then we have for the Vandermonde matrix* $\boldsymbol{A} \in \mathbb{C}^{M \times N^d}$

$$(N - \gamma)^{\frac{d}{2}} \leq N^{\frac{d}{2}} \left(\frac{\lfloor N/\gamma \rfloor}{N/\gamma}\right)^{\frac{d}{2}} \leq \sigma_{\min}(\boldsymbol{A}) \leq N^{\frac{d}{2}} \leq \sigma_{\max}(\boldsymbol{A}) \leq N^{\frac{d}{2}} \left(\frac{\lfloor N/\gamma \rfloor}{N/\gamma}\right)^{\frac{d}{2}} \leq (N + \gamma)^{\frac{d}{2}}.$$

Proof. By Remark 3.1.12 we can go to subsets of equispaced nodes in Theorem 3.5.3 and upper bounds on the largest and lower bounds on the smallest singular value stay valid. □

The respective first non-trivial bounds for the singular values in the above corollary suggest that if $\gamma = N$ both singular values equal $N^{d/2}$ and thus the Vandermonde matrix is perfectly conditioned. The following theorem shows that this is true. Furthermore, analogously to Theorem 3.1.11 we can fully characterize when the multivariate Vandermonde matrix is perfectly conditioned.

Theorem 3.5.5 (Perfectly conditioned Vandermonde matrix)**.**
Let $M, N, d \in \mathbb{N}_+$ and $\boldsymbol{A} \in \mathbb{C}^{M \times N^d}$ be a Vandermonde matrix as in (3.4.1)*. Then we have*

$$\operatorname{cond}(\boldsymbol{A}) = 1$$

if and only if for each pair of distinct nodes $\boldsymbol{t}, \boldsymbol{t}'$ there exists a component $1 \leq s \leq d$ such that $(\boldsymbol{t} - \boldsymbol{t}')_s \in \mathbb{Z}/N \setminus \{0\}$, see Figure 3.5.2 for two examples.

Proof. Analogously to the proof of Theorem 3.1.11, due to Lemma 2.1.27 we have $\operatorname{cond}(\boldsymbol{A}) = 1$ if and only if $\boldsymbol{A}\boldsymbol{A}^* = N^d \boldsymbol{I}_N$. The off diagonal entries and hence $d_{N-1}(\boldsymbol{t} - \boldsymbol{t}')$ are all zero if and only if for any distinct pair of nodes $\boldsymbol{t}, \boldsymbol{t}'$ there is one component of $\boldsymbol{t} - \boldsymbol{t}'$ being in $\mathbb{Z}/N \setminus \{0\}$. Here we used that the tensor product is zero if and only if one factor $d_{N-1}(t) = 0$ if and only if $t \in \mathbb{Z}/N \setminus 0$. □

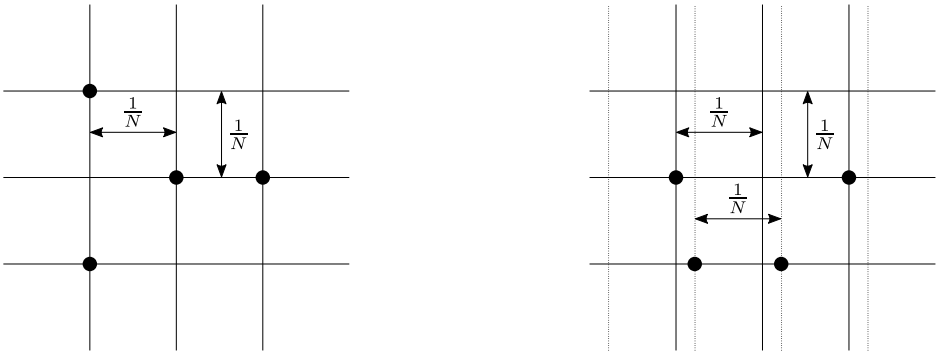


Figure 3.5.2: examples of node sets with perfectly conditioned Vandermonde matrix for $d = 2$; left: nodes on a grid with width $1/N$; right: each two distinct nodes have one coordinate direction with separation $k/N, 0 \neq k \in \mathbb{Z}$.

There is also work in connection with discrete Ingham inequalities, that are directly related to the bounds for the extremal singular values of the Vandermonde matrix, see also (3.1.11).

Lemma 3.5.6 ([66, Cor. 2.5])**.**
Let $N = 2n + 1 \in \mathbb{N}_+$. If the node set of the multivariate Vandermonde matrix $\boldsymbol{A} \in \mathbb{C}^{M \times N^d}$ from (3.4.1) *satisfies the separation condition $qn \geq 3 + 2\log(d)$ then $\sigma_{\min}(\boldsymbol{A}) > 0$. Therefore, the condition number of $\boldsymbol{A}$ is bounded.*

Remark 3.5.7.
Kunis et al. proved the logarithmic dependency on the dimension in Lemma 3.5.6. Already in [92, Cor. 3.3] it is proven that $\sigma_{\min}(\boldsymbol{A}) > 0$ but under the stronger assumption on the minimal separation distance $q(N-1) \geq 2\pi\sqrt{d}$, depending on the square root of the dimension. This result relies on the Ingham inequality from [64] and translates the Ingham inequality into bounds for singular values of $\boldsymbol{A}$.

These bounds for the smallest singular value are of qualitative nature only. Explicit bounds that allow to draw conclusions about the quantitative behavior of the condition number are not given. In fact, for proving Lemma 3.5.6 an Ingham inequality was established by means of certain localized function ψ in a similar fashion as done in the proof of Theorem 3.1.7. The function $\psi\colon \mathbb{T}^d \to \mathbb{R}$ has the properties

$$\begin{aligned}
&\widehat{\psi}(\boldsymbol{x}) \begin{cases} \geq 0, \text{ if } \|\boldsymbol{x}\|_p \leq n, \\ \leq 0, \text{ if } \|\boldsymbol{x}\|_p \geq n, \end{cases} \\
&\psi(0) > 0, \text{ if } nq > C_p \sqrt[p]{d}, \text{ where } C_p \leq \frac{2p+3}{\mathrm{e}\pi}, \\
&\text{and } \operatorname{supp}(\psi) = [-q, q]^d.
\end{aligned}$$

For $p \geq 1$, $\|\boldsymbol{x}\|_p := \left(\sum_{j=1}^m ((\boldsymbol{x})_j)^p\right)^{\frac{1}{p}}, \boldsymbol{x} \in \mathbb{C}^m$ is the well-known p-norm for vectors. The established Ingham inequality then is: for any $\boldsymbol{v} = (v_1, \ldots, v_M) \in \mathbb{C}^M$ we have

$$\max_{\boldsymbol{x}\in\mathbb{R}^d} \widehat{\psi}(\boldsymbol{x}) \sum_{\substack{\boldsymbol{\nu}\in\mathbb{Z}^d \\ \|\boldsymbol{\nu}\|_p \leq n}} \left| \sum_{j=1}^M v_j \mathrm{e}^{2\pi i \boldsymbol{\nu}^* \boldsymbol{t_j}} \right|^2 \geq \psi(0) \sum_{j=1}^M |v_j|^2 .$$

Under the assumptions of Lemma 3.5.6 with the choice $p = 2\lceil \log(d) \rceil$ the bound

$$\sigma_{\min}(\boldsymbol{A}) \geq \left(\frac{\psi(0)}{\max_{\boldsymbol{x}\in\mathbb{R}^d} \widehat{\psi}(\boldsymbol{x})} \right)^{\frac{1}{2}} > 0$$

is finally derived. Recently in [107] Anna Strotmann carefully bounded the values of $\psi(0)$ and $\max_{\boldsymbol{x}\in\mathbb{R}^d} \widehat{\psi}(\boldsymbol{x})$ and thus, extended the result to the following.

Lemma 3.5.8 ([107])**.**
Let $N \in \mathbb{N}_+$ be odd. If the node set of the multivariate Vandermonde matrix $\boldsymbol{A} \in \mathbb{C}^{M\times N^d}$ from (3.4.1) *satisfies the separation condition $q(N-1) > 3 + 2\log(d)$ then we have*

$$\sigma_{\min}(\boldsymbol{A}) \geq \frac{3}{\sqrt{10}} \left(\frac{17\sqrt{2\log(d)+4}}{20\sqrt{\pi} q} \right)^{\frac{d}{2}} > 0.$$

Unfortunately, this bound decreases with larger separation distance q and furthermore, since for instance Theorem 3.5.3 suggest that $\sigma_{\min}(\boldsymbol{A}) \approx N^{d/2}$, we need an assumption on q not being too large to make this bound effective. For example, with the additional precondition $qN \leq 2(3 + 2\log(d))$, we thus obtain

$$\sigma_{\min}(\boldsymbol{A}) \geq N^{d/2} \frac{3}{\sqrt{10}} \left(\frac{17\sqrt{2\log(d)+4}}{20\sqrt{\pi}(6+4\log(d))} \right)^{\frac{d}{2}}. \tag{3.5.1}$$

Switching the perspective, the new assumption on q limits how large N can be if q is fixed. Indeed, by Lemma 2.1.18, N can be increased and the bound stays valid, but has the initial $\widetilde{N}$ in the singular value bound. Thus the bound becomes more and more inaccurate with increasing N. To treat this problem we state the following lemma.

Lemma 3.5.9.
Let $a \in \mathbb{N}$, $d, M, N \in \mathbb{N}_+$ with $M \leq N^d$ and $\boldsymbol{A}_N \in C^{M \times N^d}$ and $\boldsymbol{A}_{aN} \in \mathbb{C}^{M \times (aN)^d}$ be Vandermonde matrices as in (3.4.1) with the same M nodes. Then it holds

$$\sigma_{\min}(\boldsymbol{A}_{aN}) \geq a^{d/2}\, \sigma_{\min}(\boldsymbol{A}_N).$$

Proof. The proof consists mainly of using the variational characterization for the smallest singular value from Theorem 2.1.15 combined with the column-wise partitioning

$$\boldsymbol{A}_{aN} = \left(\boldsymbol{D}_0 \boldsymbol{A}_N \mid \boldsymbol{D}_1 \boldsymbol{A}_N \mid \cdots \mid \boldsymbol{D}_{a^d-1} \boldsymbol{A}_N\right),$$

where the matrices $\boldsymbol{D}_k$ are defined as follows. Let $k = 0, \ldots, a^d - 1$ be an enumeration of the elements in $\{0, a-1\}^d$. Then, if k belongs to $(v_1, \ldots, v_d) \in \{0, a-1\}^d$, we set

$$\boldsymbol{D}_k = \prod_{\ell=1}^{d} \operatorname{diag}\left((\boldsymbol{z}_1)_\ell^{v_\ell N}, \ldots, (\boldsymbol{z}_M)_\ell^{v_\ell N}\right) \in \mathbb{C}^{M \times M}.$$

Each $\boldsymbol{D}_k$ is a unitary matrix and thus, they describe bijective, isometric linear maps. Therefore, we obtain

$$\begin{aligned}
\sigma_{\min}(\boldsymbol{A}_{aN})^2 &= \min_{\boldsymbol{v} \in \mathbb{C}^M,\, \|\boldsymbol{v}\|=1} \|\boldsymbol{A}_{aN}^* \boldsymbol{v}\|^2 = \min_{\boldsymbol{v} \in \mathbb{C}^M,\, \|\boldsymbol{v}\|=1} \sum_{k=0}^{a^d-1} \|\boldsymbol{A}_N^* \boldsymbol{D}_k^* \boldsymbol{v}\|^2 \\
&\geq \sum_{k=0}^{a^d-1} \min_{\boldsymbol{v} \in \mathbb{C}^M,\, \|\boldsymbol{v}\|=1} \|\boldsymbol{A}_N^* \boldsymbol{D}_k^* \boldsymbol{v}\|^2 = a^d \min_{\tilde{\boldsymbol{v}} \in \mathbb{C}^M,\, \|\tilde{\boldsymbol{v}}\|=1} \|\boldsymbol{A}_N^* \tilde{\boldsymbol{v}}\|^2 = a^d\, \sigma_{\min}(\boldsymbol{A}_N)^2.
\end{aligned}$$

□

With this lemma at hand we can utilize the bound from Lemma 3.5.8 for arbitrary large orders N. Though, we have to pay with a slightly worse constant in the bound.

Theorem 3.5.10.
Let $N \in \mathbb{N}_+$ and $\boldsymbol{A} \in \mathbb{C}^{M \times N^d}$ be a Vandermonde matrix as in (3.4.1) having M well-separated nodes with separation distance satisfying

$$q(N-1) > 3 + 2\log(d).$$

Then we have for the smallest singular value

$$\sigma_{\min}(\boldsymbol{A}) \geq N^{d/2} \frac{3}{\sqrt{10}} \left(\frac{17\sqrt{2\log(d)+4}}{60\sqrt{\pi}(3+2\log(d))} \right)^{\frac{d}{2}}.$$

Proof. We set $c_d := 3 + 2\log(d)$. If $N \leq \frac{3c_d}{2q}$, we can reorder this to an upper bound for q and obtain with Lemma 3.5.8

$$\sigma_{\min}(\boldsymbol{A}) \geq N^{d/2} \frac{3}{\sqrt{10}} \left(\frac{17\sqrt{2\log(d)+4}}{30\sqrt{\pi}(3+2\log(d))} \right)^{\frac{d}{2}}$$

from which the bound in the statement of the theorem follows. If $N > \frac{3c_d}{2q}$, we define the odd number

$$\widetilde{N} := 2\left\lfloor \frac{c_d}{2q} + 1 \right\rfloor + 1$$

that fulfills

$$\frac{c_d}{\widetilde{N}-1} < \frac{c_d}{c_d/q} = q$$

and, since $q \leq 1/2$ and thus $6 \leq c_d/q$, also

$$\frac{3c_d}{2\widetilde{N}} \geq \frac{3c_d}{2c_d/q+6} \geq q. \tag{3.5.2}$$

Now we can write $N = a\widetilde{N} + r$ for some $a \in \mathbb{N}_+$ and $r < \widetilde{N}$. Furthermore, it holds $N = a\widetilde{N} + r \leq 2a\widetilde{N}$ (the constant 2 is best possible for a bound of the form $N \leq ca\widetilde{N}, c \in \mathbb{R}_{>0}$, see the case $a = 1$). Using additionally Lemmata 2.1.18, 3.5.8 and 3.5.9 and (3.5.2) we finally obtain

$$\begin{aligned}\sigma_{\min}(\boldsymbol{A}) = \sigma_{\min}(\boldsymbol{A}_N) \geq \sigma_{\min}(\boldsymbol{A}_{a\widetilde{N}}) \geq a^{d/2}\,\sigma_{\min}(\boldsymbol{A}_{\widetilde{N}}) &\geq a^{d/2}\frac{3}{\sqrt{10}}\left(\frac{17\sqrt{2\log(d)+4}}{20\sqrt{\pi}q}\right)^{\frac{d}{2}} \\ &\geq (a\widetilde{N})^{d/2}\frac{3}{\sqrt{10}}\left(\frac{17\sqrt{2\log(d)+4}}{30\sqrt{\pi}(3+2\log(d))}\right)^{\frac{d}{2}} \geq N^{d/2}\frac{3}{\sqrt{10}}\left(\frac{17\sqrt{2\log(d)+4}}{60\sqrt{\pi}(3+2\log(d))}\right)^{\frac{d}{2}}.\end{aligned}$$

□

In [75, Thm. 1] an upper bound on the largest singular value in the case of coordinate-wise well-separated node sets is given, i.e. for node sets with $q_s > \frac{1}{N}$, where q_s is the minimal wrap around distance in the s-th coordinate direction for $s = 1, \ldots, d$. In fact the result is stated also for different orders N_s in each direction, but we simplify here a little. The bound is then given by

$$\sigma_{\max}(\boldsymbol{A}) \leq \prod_{s=1}^{d}\left(N - 1 + \frac{1}{q_s}\right). \tag{3.5.3}$$

A problem arises in our setting since if we have well-separated nodes as in Definition 3.5.1 the bound becomes infinity already for nodes that are on one coordinate axis and in that direction well-separated. The proof in [75, Appendix A] uses the tensor product of the univariate Selberg majorant. We adapt it here to establish an upper bound for the case of our more general definition of well-separated nodes.

Lemma 3.5.11 (Bounds on the largest singular value)**.**
Let $d, N \in \mathbb{N}_+$ and $\boldsymbol{A}$ be a multivariate Vandermonde matrix as in (3.4.1) *with minimal separation distance q. Then the largest singular value of $\boldsymbol{A}$ is bounded from above by*

$$N^{\frac{d}{2}} \leq \sigma_{\max}(\boldsymbol{A}) \leq \left(N - 1 + \frac{1}{q}\right)^{\frac{d}{2}} \leq N^{\frac{d}{2}}\left(1 + \frac{1}{qN}\right)^{\frac{d}{2}}$$

independent of the number of nodes.

Proof. We proceed similar to the proof of Theorem 3.1.7 and use the Selberg majorant u from Lemma 2.4.3 for the characteristic function on the interval $[0, N-1]$ This time we build its d-fold tensor product to obtain

$$g\colon \mathbb{R}^d \to \mathbb{R}, \quad g(\boldsymbol{x}) = g(x_1, \ldots, x_d) := \prod_{s=1}^{d} u(x_s),$$

which is a majorant for $\chi^d_{[0,N-1]}$. By Fubini's theorem its multivariate Fourier transform is again given by the tensor product of the univariate Fourier transform,

$$\widehat{g}\colon \mathbb{R}^d \to \mathbb{R}, \quad \widehat{g}(\boldsymbol{\omega}) = \widehat{g}(\omega_1, \ldots, \omega_d) := \prod_{s=1}^{d} \widehat{u}(\omega_s),$$

and hence, $\widehat{g}$ is supported in $[-q, q]^d$. Now we can bound the norm of our Vandermonde matrix. We use the variational characterization of the largest singular value from Theorem 2.1.15. For arbitrary $\boldsymbol{v} \in \mathbb{C}^M$ we have

$$\begin{aligned}
\|\boldsymbol{A}^*\boldsymbol{v}\|^2 &= \sum_{\boldsymbol{\nu}\in\mathbb{N}^d,\, \|\boldsymbol{\nu}\|_\infty < N} \left| \sum_{j=1}^{M} v_j \mathrm{e}^{-2\pi i \boldsymbol{\nu}^* \boldsymbol{t}_j} \right|^2 = \sum_{\boldsymbol{\nu}\in\mathbb{Z}^d} \chi_{[0,N-1]^d}(\boldsymbol{\nu}) \left| \sum_{j=1}^{M} v_j \mathrm{e}^{-2\pi i \boldsymbol{\nu}^* \boldsymbol{t}_j} \right|^2 \\
&\leq \sum_{\boldsymbol{\nu}\in\mathbb{Z}^d} g(\boldsymbol{\nu}) \left| \sum_{j=1}^{M} v_j \mathrm{e}^{-2\pi i \boldsymbol{\nu}^* \boldsymbol{t}_j} \right|^2 = \sum_{\boldsymbol{\nu}\in\mathbb{Z}^d} g(\boldsymbol{\nu}) \left(\sum_{j=1}^{M} \overline{v}_j \mathrm{e}^{2\pi i \boldsymbol{\nu}^* \boldsymbol{t}_j} \right) \left(\sum_{k=1}^{M} v_k \mathrm{e}^{-2\pi i \boldsymbol{\nu}^* \boldsymbol{t}_k} \right) \\
&= \sum_{j=1}^{M} \sum_{k=1}^{M} \overline{v}_j v_k \sum_{\boldsymbol{\nu}\in\mathbb{Z}^d} g(\boldsymbol{\nu}) \mathrm{e}^{-2\pi i \boldsymbol{\nu}^* (\boldsymbol{t}_k - \boldsymbol{t}_j)} = \sum_{j=1}^{M} \sum_{k=1}^{M} \overline{v}_j v_k \sum_{\boldsymbol{r}\in\mathbb{Z}^d} \widehat{g}(\boldsymbol{t}_k - \boldsymbol{t}_j + \boldsymbol{r}),
\end{aligned}$$

where the last equality follows from the Poisson summation formula. The support of g is in $[-q, q]^d$ and since the minimal separation distance of the nodes $\boldsymbol{t}_1, \ldots, \boldsymbol{t}_M$ is q, we have $\widehat{g}(\boldsymbol{t}_k - \boldsymbol{t}_j + \boldsymbol{r}) = 0$ if $j \neq k$. Notice that for every pair of nodes $\boldsymbol{t}_j, \boldsymbol{t}_k, j \neq k$ there is one coordinate direction s for which $\left|(\boldsymbol{t}_k - \boldsymbol{t}_j)_s\right|_{\mathbb{T}} \geq q$. Therefore, the corresponding component of the tensor product, $\widehat{g}$ is made of, is zero no matter what value $(\boldsymbol{r})_s$ has if $j \neq k$. Proceeding the above calculation and using Lemma 2.4.3 leads to

$$\|\boldsymbol{A}^*\boldsymbol{v}\|^2 \leq \|\boldsymbol{v}\|^2 \, \widehat{g}(\boldsymbol{0}) = \|\boldsymbol{v}\|^2 \prod_{s=1}^{d} \widehat{u}(0) = \|\boldsymbol{v}\|^2 \left(N - 1 + \frac{1}{q} \right)^d.$$

Finally, dividing by $\|\boldsymbol{v}\|$, taking the square root and using the variational characterization of the matrix norm yields the result, since $\boldsymbol{v}$ was arbitrary. □

Remark 3.5.12.
In [75] also the tensor product of a univariate minorant is used to establish a lower bound on the smallest singular value in the same fashion as in (3.5.3)*. Unfortunately, the tensor product of the minorant for the characteristic function on an interval is assumed to be a minorant for the characteristic function of the multivariate box, but already in the bivariate case, there are areas on the plane outside the square, where the tensor product is positive since both factors of the tensor product are negative. Hence, the tensor product is not a minorant anymore and the argumentation is wrong.*

Theorem 3.5.13 (Bounds on the smallest singular value)**.**
Let $\boldsymbol{A} \in \mathbb{C}^{M \times N^d}$ be a multivariate Vandermonde matrix of well-separated nodes with normalized minimal separation distance satisfying

$$qN \geq 2(d+1).$$

Then for the smallest singular value of $\boldsymbol{A}$ we have

$$\sigma_{\min}(\boldsymbol{A}) \geq N^{\frac{d}{2}} \left(\frac{1}{(d+1)^d} - \frac{\pi^2 (2^d - 1)(d+1)}{12(qN)^{d+1}} \right)^{1/2} > N^{\frac{d}{2}} \frac{3}{4(d+1)^{d/2}}.$$

If there are two nodes in Ω with separation distance q in each coordinate direction, then we have

$$\sigma_{\min}(\boldsymbol{A}) \leq N^{\frac{d}{2}} \left(1 - \frac{1}{(\pi q N)^d} \right)^{1/2}.$$

Proof. We adapt the technique from Section 3.3.4 and apply Lemma 2.1.17 with appropriate diagonal matrix $\boldsymbol{C}$, such that each entry of $\boldsymbol{ACA}^*$ becomes the evaluation of a trigonometric polynomial with specific decay property. Afterwards, the Gerschgorin disc theorem is applied to lower bound the smallest singular value, cf. [71, Thm. 4.6].

Let $\beta \in \mathbb{N}_+$, $\beta \geq d+1$ and $m := \left\lfloor \frac{N-1}{\beta} \right\rfloor$. Hence, the β-th power of the univariate modified Dirichlet kernel of degree m, d_m^β, see Definition 2.4.10, has degree $\beta m \leq N-1$ and we have $(m+1)\beta \geq N$. The latter can be seen by writing $N - 1 = a\beta + b$, where $a, b \in \mathbb{N}$ and $b < \beta$. Therefore, $m = a$ and

$$(m+1)\beta = a\beta + \beta > a\beta + b = N - 1. \tag{3.5.4}$$

We define the diagonal matrix

$$\boldsymbol{C} := \operatorname{diag}\left((m+1)^d \widehat{(d_m^\beta)}(\boldsymbol{\nu}) \right)_{\substack{\boldsymbol{\nu} \in \mathbb{N}^d \\ \|\boldsymbol{\nu}\|_\infty < N}} \tag{3.5.5}$$

with diagonal entries ordered according to the columns of $\boldsymbol{A}$. Notice that these entries consist of all non-zero Fourier coefficients of $(m+1)^d d_m^\beta$ with possibly additional zeros depending on $m\beta$ being smaller or equal to $N-1$. Lemma 2.4.11 v) guarantees that $(m+1)^d \widehat{(d_m^\beta)}(\boldsymbol{\nu}) \leq 1$ and hence, $\|\boldsymbol{C}\| \leq 1$, which is therefore a Hermitian contraction. This enables us to apply Lemma 2.1.17 and we get

$$\lambda_{\min}(\boldsymbol{AA^*}) \geq \lambda_{\min}(\boldsymbol{ACA}^*),$$

where the entries of the matrix on the right hand side are given by

$$\left(\boldsymbol{ACA}^*\right)_{(j,k)} = \sum_{\boldsymbol{\nu} \in \mathbb{N}^d,\, \|\boldsymbol{\nu}\|_\infty < N} \mathrm{e}^{2\pi i \boldsymbol{\nu}^* \boldsymbol{t}_j} (m+1)^d \widehat{(d_m^\beta)}(\boldsymbol{\nu}) \mathrm{e}^{-2\pi i \boldsymbol{\nu}^* \boldsymbol{t}_k} = (m+1)^d d_m^\beta(\boldsymbol{t}_j - \boldsymbol{t}_k).$$

Lemma 2.4.11 ii) shows that $\left|(m+1) d_m^\beta(\boldsymbol{t})\right| \leq \frac{1}{2^\beta (m+1)^\beta |\boldsymbol{t}|_{\mathbb{T}^d}^\beta}$. Together with Lemma 2.1.22, $d_m^\beta(0) = 1$ from Lemma 2.4.11 i), the packing argument from Lemma 3.4.2 and $\beta(m+1) \geq N$ from above this yields

$$\lambda_{\min}(\boldsymbol{ACA}^*) \geq \min_{1 \leq j \leq M} \left\{ (m+1)^d - (m+1)^d \sum_{k=1,\, k \neq j}^{M} \left| d_m^\beta(\boldsymbol{t}_j - \boldsymbol{t}_k) \right| \right\}$$

$$
\begin{aligned}
&\geq (m+1)^d - (m+1)^d \max_{1\leq j\leq M} \left\{ \sum_{k=1,\, k\neq j}^{M} \frac{1}{2^\beta (m+1)^\beta \, |\boldsymbol{t}_j - \boldsymbol{t}_k|_{\mathbb{T}^d}^\beta} \right\} \\
&\geq (m+1)^d - (m+1)^d \sum_{\ell=1}^{\left\lfloor \frac{1}{2q} \right\rfloor} \frac{2^d (2^d-1)\ell^{d-1}}{2^\beta (m+1)^\beta} \max_{\boldsymbol{t}\in J_\ell} \left\{ \frac{1}{|\boldsymbol{t}|_{\mathbb{T}^d}^\beta} \right\} \\
&\geq (m+1)^d - (m+1)^d \frac{2^d(2^d-1)}{2^\beta (m+1)^\beta q^\beta} \sum_{\ell=1}^{\left\lfloor \frac{1}{2q} \right\rfloor} \frac{\ell^{d-1}}{\ell^\beta} \\
&\geq (m+1)^d \left(1 - \frac{(2^d-1)\zeta(\beta-d+1)}{2^{\beta-d}(m+1)^\beta q^\beta} \right) \\
&= (\beta(m+1))^d \left(\frac{1}{\beta^d} - \frac{(2^d-1)\zeta(\beta-d+1)\beta^\beta}{2^{\beta-d}\beta^d \left(\beta(m+1)\right)^\beta q^\beta} \right) \\
&\geq N^d \left(\frac{1}{\beta^d} - \frac{(2^d-1)\zeta(\beta-d+1)\beta^\beta}{2^{\beta-d}\beta^d (qN)^\beta} \right)
\end{aligned}
$$

where ζ denotes the Riemann zeta function. If we choose $\beta = d+1$, then $\zeta(\beta+d-1) = \zeta(2) = \frac{\pi^2}{6}$ and we finally obtain

$$
\lambda_{\min}(\boldsymbol{A}\boldsymbol{A}^*) \geq N^d \left(\frac{1}{(d+1)^d} - \frac{\pi^2 (2^d-1)(d+1)}{12(qN)^{d+1}} \right). \tag{3.5.6}
$$

Now the first bound for $\boldsymbol{A}$ directly follows by Lemma 2.1.12. Using the assumptions on qN from the statement of the theorem in (3.5.6), we obtain

$$
\sigma_{\min}(\boldsymbol{A}) \geq N^{d/2} \frac{1}{(d+1)^{d/2}} \left(1 - \frac{\pi^2(2^d-1)}{12 \cdot 2^{d+1}} \right)^{\frac{1}{2}} \geq N^{d/2} \frac{1}{(d+1)^{d/2}} \sqrt{1 - \frac{\pi^2}{24}} \geq N^{d/2} \frac{3}{4(d+1)^{d/2}},
$$

which is the second bound.

The upper bound follows similarly to the univariate case, see Theorem 3.1.10. Assume there are two nodes $\boldsymbol{t}, \boldsymbol{t}'$ that realize the minimal separation distance $q = \frac{k+1}{2N}, k \in \mathbb{N}_+$ in each coordinate. Applying Lemma 2.1.6 to $\boldsymbol{A}\boldsymbol{A}^*$ yields

$$
\sigma_{\min}(\boldsymbol{A})^2 \leq N^d \, \sigma_{\min} \left(\begin{pmatrix} 1 & d_{N-1}(\boldsymbol{t}-\boldsymbol{t}') \\ d_{N-1}(\boldsymbol{t}'-\boldsymbol{t}) & 1 \end{pmatrix} \right) = N^d \left(1 - \left| d_{N-1}(\boldsymbol{t}-\boldsymbol{t}') \right| \right)
$$

with modified Dirichlet kernel d_{N-1} from Lemma 2.4.11. Using Definition 2.4.10, the identity $\left| \sin(\frac{k+1/2}{N} N\pi) \right| = 1$ and $\sin(\pi q) \leq q\pi$ concludes the second part. □

Remark 3.5.14.
Choosing $\boldsymbol{C}$ such that simply powers of modified Dirichlet kernels are created via $\boldsymbol{A}\boldsymbol{C}\boldsymbol{A}^*$ without the prefactor $(m+1)^d$ would lead to a missing normalization factor in the end. Irrespective of that, for $d = 1$ and $qN > \frac{\pi}{\sqrt{3}} \approx 1.814$ we obtain

$$
\sigma_{\min}(\boldsymbol{A}) \geq \frac{1}{\sqrt{2}} \sqrt{N} \sqrt{1 - \frac{\pi^2}{3(qN)^2}},
$$

which is weaker than the bound from Theorem 3.1.7.

Finally, we use quite general results for clustered node configurations. Since Theorem 3.4.12 provides lower bounds for the smallest singular value of Vandermonde matrices with clustered node configurations of arbitrary cluster sizes, together with Remark 3.5.2 we can readily derive a bound for the case of well-separated node sets.

Corollary 3.5.15 (Lower bound on the smallest singular value)**.**
Let $M, d, N \in \mathbb{N}_+$, $N > \max\left\{M, 2(d+2)^2\right\}$ and $\boldsymbol{A} \in \mathbb{C}^{M \times N^d}$ be a Vandermonde matrix as in (3.4.1) *having M well-separated nodes with separation distance satisfying*

$$qN > 6d.$$

Then from Theorem 3.4.12 with $\lambda = 1$ we obtain

$$\sigma_{\min}(\boldsymbol{A}) \geq \frac{N^{d/2}}{3d^{d/4}}.$$

Note that Theorem 3.4.12 always assumes $qN \geq \beta \geq (d+1)$. If $d = 1$ and $N > \max\{M, 8\}$, then $qN \geq 4.4$ implies

$$\sigma_{\min}(\boldsymbol{A}) \geq \frac{\sqrt{N}}{1.8}.$$

This compares to Theorem 3.1.10, which provides $\sigma_{\min}(\boldsymbol{A}) \geq \sqrt{N} \cdot \sqrt{1 - 1/(qN)} \geq \sqrt{N}/1.14$ under the same condition on q.

Proof. Since $\lambda = 1$, we have $\mathcal{C} = 1$. Corollary 3.4.14 i) yields the first result.

If $d = 1$, then applying Theorem 3.4.12 with $\beta = 2$ and $\zeta(2) = \pi^2/6$ needs the condition

$$qN \geq \frac{2 \cdot (2^{1/4}) \cdot \sqrt{2}\pi}{\sqrt{6}},$$

where the right hand side is smaller than 4.4, and yields

$$\sigma_{\min}(\boldsymbol{A}) \geq (1.5 \cdot 2^{1/4})^{-1}\sqrt{N} \geq \frac{\sqrt{N}}{1.8}.$$

□

Now we have several results that state lower bounds for the smallest singular value of multivariate Vandermonde matrices with well-separated nodes. We conclude by providing a table that simplifies the comparison between the results, Table 3.5.1.

	$q(N-1) \gtrsim$	$\sigma_{\min}(\boldsymbol{A}) \gtrsim$
[92, Cor. 3.3]	$\sqrt{d}$	is positive
Theorem 3.5.10	$\log(d)$	$\left(\frac{c}{\log(d)}\right)^{\frac{d}{4}} N^{\frac{d}{2}}$
Theorem 3.5.13	d	$\left(\frac{1}{d+1}\right)^{d} N^{\frac{d}{2}}$
Corollary 3.5.15	d	$\left(\frac{1}{d}\right)^{\frac{d}{4}} N^{\frac{d}{2}}$

Table 3.5.1: qualitative bounds for the smallest singular value of multivariate Vandermonde matrices with well-separated nodes showing the dependency on the spatial dimension d ($\gtrsim$ means greater than, up to a constant independent on quantities on the right hand side).

Chapter 4

Stability of the ESPRIT method

This chapter deals with an application of the results obtained from studying the condition number of rectangular Vandermonde matrices in Chapter 3. We look at the problem of reconstructing parameters of an exponential sum. One class of solution algorithms are the subspace methods and we concentrate on one of these methods, namely the estimation of parameters via rotational invariance techniques (ESPRIT) and refer to it as ESPRIT method or ESPRIT algorithm. We show that the stability of this method, i.e. the error amplification that occur when applied to corrupted data, can be bounded in terms of the condition number of involved Vandermonde matrices.

4.1 Reconstruction of exponential sums

We start with the definition of an exponential sum and introduce related notation. Let $N, M \in \mathbb{N}$, $\Omega = \{t_1, \dots, t_M\} \subset \mathbb{T}$ distinct nodes with representation $z_j = \mathrm{e}^{2\pi i t_j} \in \mathbb{C}, j = 1, \dots, M$, on the complex unit circle collected in the set $\Lambda := \{z_1, \dots, z_M\}$. Let $\alpha_1, \dots, \alpha_M \in \mathbb{C} \setminus \{0\}$ coefficients, then the *exponential sum* $\mathcal{E}\colon \mathbb{Z} \to \mathbb{C}$ of *order* M is given by

$$\mathcal{E}(k) := \sum_{j=1}^{M} \alpha_j z_j^k. \tag{4.1.1}$$

We assume that N samples $\mathcal{E}_k := \mathcal{E}(k), k = 0, \dots, N-1$, and the order M are given. More practically relevant is the situation in which we have perturbed samples

$$\widetilde{\mathcal{E}}_k := \mathcal{E}(k) + e_k, \quad k = 0, \dots, N-1, \tag{4.1.2}$$

with noise vector $\boldsymbol{e} = (e_0, \dots, e_{N-1})$. The noise can occur through measurement errors, or rounding errors when storing the data. Then the task is to reconstruct the nodes and coefficients using the given information about $\mathcal{E}$.

Exponential sums and the task to reconstruct its parameters appear in a variety of applications and scientific fields, e.g. radar and sonar [25], direction of arrival estimations [65, 99], exponential data fitting [20, 86], time series analysis [105] and super-resolution microscopy [37]. For more details, we also refer to the survey [90], the manuscript [22], and references therein.

The unperturbed problem, with only a few summands, was already tackled by Prony in the 18th century, see [30]. As modern research shows, his solution technique can be applied

for the above described problem [90] and even in more general [87, 67, 61] and multivariate settings [92, 70, 98, 28]. We briefly describe the so called Prony's method for solving the unperturbed problem. The first observation is that the task can be split into two parts. Obviously, once the nodes z_j are reconstructed the coefficients can be calculated by solving an (over-determined) linear system of equations. In order to reconstruct the nodes Prony discovered that, assuming the nodes are known, one can set up the complex polynomial

$$p(z) = \prod_{j=1}^{M} (z - z_j)$$

which vanishes at each node and only there. This polynomial is nowadays called *Prony polynomial*. It can be represented in the monomial basis

$$p(z) = \sum_{k=0}^{M} c_k z^k,$$

with coefficients $c_k \in \mathbb{C}$. Combining these with an array of samples, we have for $\ell \in \mathbb{N}$ the equation

$$\sum_{k=0}^{M} c_k \mathcal{E}_{k+\ell} = \sum_{k=0}^{M} c_k \sum_{j=1}^{M} \alpha_j z_j^{k+\ell} = \sum_{j=1}^{M} z_j^{\ell} \sum_{k=0}^{M} c_k z_j^k = \sum_{j=1}^{M} z_j^{\ell} p(z_j) = 0.$$

Since by definition of the Prony polynomial as a product over linear terms with slope one, it is monic, i.e. the leading coefficient c_M equals one. Therefore, the equations with $\ell = 0, \ldots, M-1$ are sufficient to solve for the coefficients of the Prony polynomial and only $N = 2M$ samples are needed. Indeed, one expects to need this number of samples, since there are $2M$ unknowns involved assuming the order of the exponential sum is a priori known. Finally, after using the calculated coefficients to set up the Prony polynomial the nodes can be reconstructed by finding its roots.

In the later 20th and early 21st centuries Prony's method reappeared in different fields. For example in context of signals with finite rate of innovation [111], the annihilating filters are in principle the coefficients of certain Prony polynomials. The pure Prony's method only works as long as the samples are unperturbed, therefore several more robust solution methods were invented to deal with noisy samples. On the one hand, there is the class of subspace or parametric methods, see [105], which include the Pisarenko method [88], a special case of the more general multiple signal classification (MUSIC) algorithm [99], the matrix pencil (MP) method and the estimation of signal parameters via rotational invariance techniques (ESPRIT) method [96]. The calculation of the roots of the Prony polynomial can be done by computing the eigenvalues of the corresponding companion matrix (see [55, Cor. 3.3.4]) and in [93] it is shown that this companion matrix is the connection between the subspace methods and Prony's method. Thus, these methods together with the approximate Prony's method (APM) [91] can also be regarded as Prony-like methods.

On the other hand, there are convex optimization approaches [23, 108, 36, 33]. The link to our situation relies on the following idea, see [23]. Instead of dealing with the exponential sum, one tries to reconstruct the related discrete complex measure $\mu : \mathbb{T} \to \mathbb{C}$,

$$\mu(t) := \sum_{j=1}^{M} \alpha_j \delta(t + t_j).$$

It's discrete Fourier data is given by

$$\hat{\mu}(k) = \int_{\mathbb{T}} \mathrm{e}^{-2\pi i k t}\, d\mu(t) = \sum_{j=1}^{M} \alpha_j \mathrm{e}^{2\pi i k t_j} = \mathcal{E}(k), \quad k \in \mathbb{Z}.$$

Therefore, reconstructing an exponential sum using its low-end integer evaluations is equivalent to reconstructing a discrete complex measure given its low end discrete Fourier data (notice that the measure is without loss of generality supported at the negative nodes due to our definition of the Fourier coefficients of a measure). In order to deal with the new formulated problem, it is again divided into two parts. The second part is the same as for exponential sums and the first part, is to reconstruct the support of the measure. This is done by finding the optimal measure in the program

$$\min_{\nu \in \mathcal{M}(\mathbb{T})} \|\nu\|_{\mathrm{TV}}, \quad \text{subject to} \quad \hat{\nu}(k) = \mathcal{E}_k, \quad k = 0, \ldots, N-1,$$

where $\mathcal{M}(\mathbb{T})$ is the set of complex measures given in Definition 2.3.5 and $\|\cdot\|_{TV}$ is the total variation norm on this space, see e.g. [97, 6.5]. Since this is an infinite dimensional problem, the dual problem is taken and lifted to a semidefinite program (SDP). The solution to this provides a polynomial that is zero at the support of the measure. The parallel to the Prony polynomial becomes apparent. A so called dual certificate ensures that the solution to SDP and therefore the dual program is also the right and unique solution to the initial program. In this case the dual certificate is the existence of a certain interpolating trigonometric polynomial. The optimization methods are proven to be stable when nodes are well-separated and the situation for clustered nodes with positive coefficients can be found in [32] and in [83, 82] for nodes on an arbitrary fine grid.

Almost all available stability analysis, also that for subspace methods, concentrate on the node reconstruction only. A good source for a stability result of the coefficient reconstruction can be found in [94], which we also recap later on in combination with our own results.

Until a few years ago, the stability analysis of the subspace methods concerning the node reconstruction was only of asymptotic or statistical nature, i.e. for $N \to \infty$ or signal to noise ratio (SNR) to infinity [106, 39], or with statistical priors like the assumption of having Gaussian distributed noise [75]. A deterministic performance analysis, i.e. the analysis of the error amplification only relying on a given level of noise, like e.g. an upper bound on the norm of the noise vector, was missing in particular for a finite number of samples N. First results in that direction were stated for the ESPRIT algorithm in [16] and [94]. Although, with the drawback that the results were of implicit nature since a solution vector of a system of equations is involved.

Both, a new variant of the matrix pencil algorithm and deterministic stability analysis for it are presented in [80], but unfortunately the proof has some inaccuracies, see [5, p. 172] and the recent manuscript [26], which provides corrections. In [6] (details of proofs can be found in [5]) an improved analysis for the most important subspace methods, including MUSIC, MP and ESPRIT, is presented. The error amplification is in terms of the absolute values of the smallest and largest coefficients of the exponential sum, the number of samples and the smallest and largest singular values of Vandermonde matrices as dealt with in Chapter 3.

Quite recently, in [13] (based on work done in [14, 2, 3, 9, 10]) the min-max error for the problem of reconstructing an exponential sum was investigated for the case that the node set consists of one cluster among several well-separated nodes. Furthermore, the setting

is slightly different since it is supposed that continuous samples $\widetilde{\mathcal{E}}(\omega) = \mathcal{E}(\omega) + e(\omega), \omega \in [-\Omega, \Omega]$ are known up to a bounded error function $e(\omega)$ with $\max_{\omega\in[-\Omega,\Omega]} e(\omega) \leq e_{\max} < \infty$. The min-max of error is of information theoretic nature and, roughly speaking, tells that each algorithm solving this problem can not perform better, but there is a certain algorithm with this performance. They proved upper and lower bounds in which the focus lies on the dependence on the minimal separation distance q, the band limit Ω and number of nodes λ forming the cluster. The result is summed up in Table 4.1.1. The question that arises is the

	outside the cluster	inside the cluster
error for reconstructing a node	$\lesssim \frac{1}{\Omega} e_{\max}$	$\lesssim \frac{1}{\Omega}(q\Omega)^{-2\lambda+2} e_{\max}$
error for reconstructing a coefficient	$\lesssim e_{\max}$	$\lesssim (q\Omega)^{-2\lambda+1} e_{\max}$

Table 4.1.1: error bounds for an optimal algorithm according to [13] under the assumption $e_{\max} \lesssim (q\Omega)^{2\lambda-1}$ on the noise.

following. Is there a practical algorithm that performs like the min-max error predicts? In [13] also numerical evidence is provided that the matrix pencil method is such an algorithm. A proof for the theoretical part is still missing in connection with the MP method.

Most recently the authors of [74] studied the stability of the ESPRIT method regarding the reconstruction of clustered nodes of an exponential sum. The error bound has the same scaling with respect to the minimal separation distance but a factor of $\frac{1}{N}$ is missing. Furthermore, reaching the correct order in qN was payed with further restrictions on the noise level. The results are summed up in table Table 4.1.2. Their analysis relies on the same technique as

noise assumption	error for reconstructing nodes
$\|\boldsymbol{e}\|_\infty \lesssim (qN)^{2\lambda-2}$	$\lesssim (qN)^{-3\lambda+3} \|\boldsymbol{e}\|_\infty$
$\|\boldsymbol{e}\|_\infty \lesssim \frac{1}{N}(qN)^{4\lambda-3}$	$\lesssim (qN)^{-2\lambda+2} \|\boldsymbol{e}\|_\infty$

Table 4.1.2: error bounds for reconstruction of nodes in clustered node configurations via the ESPRIT method by [74].

used in [5], with the difference that they make use of principal angles and vectors allowing a more compact proof. Furthermore, they refined the analysis of the smallest singular value of a certain truncated matrix.

The purpose of this chapter is, despite of optimizing the presentation of the stability analysis by [74] of the ESPRIT algorithm, to show that it basically provides the same result as [6] until a certain point. In [74] improved estimates with respect to the dependence on the minimal separation distance are provided, but an additional dependence on the number of nodes M is obtained. We show that the latter is artificial by a more detailed analysis of principal vector matrices. For practical use of the stability bounds, we show that staying with condition bounds for Vandermonde matrices instead of using a so called uncertainty principle estimate from [74] is in particular more appropriate for a large number of nodes in the well-separated case and for small clusters compared to the total number of nodes in the case where nodes build clusters.

First of all, we recapitulate the ESPRIT method in the next section. Afterwards we present a stability analysis of its node reconstruction in which different bounds from [5, 74] for the smallest singular value of an involved truncated matrix are compared. Finally, a stability analysis of ESPRIT for reconstructing the coefficients is presented.

4.2 ESPRIT algorithm

The ESPRIT algorithm is based on the following observations. We start with the unperturbed problem (4.1.1). Since we are dealing with with Vandermonde matrices of different degrees but with the same node sets, we keep the index corresponding to the degree. For $L \in \mathbb{N}_+$ set

$$\boldsymbol{A}_L = \left(z_j^{k-1}\right)_{j=1,k=1}^{M,L} = \begin{pmatrix} 1 & z_1 & \cdots & z_1^{L-1} \\ \vdots & \vdots & \vdots & \vdots \\ 1 & z_M & \cdots & z_M^{L-1} \end{pmatrix} \in \mathbb{C}^{M\times L}. \tag{4.2.1}$$

We assume that $M < L \leq N - M + 1$, such that this matrix has full rank M, since all nodes are distinct. Corresponding to the given samples, we define the Hankel matrix (sometimes also called L-trajectory matrix [94])

$$\boldsymbol{H}_L := \left(\mathcal{E}_{k+\ell}\right)_{k=0,\ell=0}^{L-1,N-L} = \begin{pmatrix} \mathcal{E}_0 & \mathcal{E}_1 & \cdots & \mathcal{E}_{N-L} \\ \mathcal{E}_1 & \mathcal{E}_2 & \cdots & \mathcal{E}_{N-L+1} \\ \vdots & \vdots & \vdots & \vdots \\ \mathcal{E}_{L-1} & \mathcal{E}_L & \cdots & \mathcal{E}_{N-1} \end{pmatrix} \in \mathbb{C}^{L\times(N-L+1)}. \tag{4.2.2}$$

The Hankel matrix satisfies the Vandermonde factorization

$$\boldsymbol{H}_L = \boldsymbol{A}_L^\top \operatorname{diag}(\alpha_1, \ldots, \alpha_M)\boldsymbol{A}_{N-L+1} \tag{4.2.3}$$

and since the involved matrices have full rank, it holds $\operatorname{rank}(\boldsymbol{H}_L) = M$ and $\operatorname{range}(\boldsymbol{H}_L) = \operatorname{range}(\boldsymbol{A}_L^\top)$. Now, we can apply the economic SVD (2.1.4) to $\boldsymbol{H}_L$ and get

$$\boldsymbol{H}_L = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^*,$$

where $\boldsymbol{U} = [\boldsymbol{u}_1, \ldots, \boldsymbol{u}_M]$ has orthonormal columns $\boldsymbol{u}_j, j = 1, \ldots, M$, and $\operatorname{range}(\boldsymbol{U}) = \operatorname{range}(\boldsymbol{A}_L^\top)$. Moreover, the columns of $\boldsymbol{U}$ and $\boldsymbol{A}_L^\top$ build bases for the same subspace of $\mathbb{C}^L$ and therefore, we find a regular matrix $\boldsymbol{P} \in \mathbb{C}^{M\times M}$, such that

$$\boldsymbol{U} = \boldsymbol{A}_L^\top \boldsymbol{P}. \tag{4.2.4}$$

We introduce the matrices

$$\boldsymbol{I}_\uparrow := \begin{pmatrix} 1 & & 0 \\ & \ddots & \vdots \\ & 1 & 0 \end{pmatrix} \in \mathbb{C}^{(L-1)\times L} \quad \text{and} \quad \boldsymbol{I}_\downarrow := \begin{pmatrix} 0 & 1 & \\ \vdots & & \ddots \\ 0 & & 1 \end{pmatrix} \in \mathbb{C}^{(L-1)\times L}, \tag{4.2.5}$$

which multiplied from left to a matrix pick its $L-1$ first rows or $L-1$ last rows. Let $\boldsymbol{U}_\uparrow := \boldsymbol{I}_\uparrow\boldsymbol{U}$, $\boldsymbol{U}_\downarrow := \boldsymbol{I}_\downarrow\boldsymbol{U}$, $\boldsymbol{A}_\uparrow := \boldsymbol{I}_\uparrow\boldsymbol{A}_L^\top$ and $\boldsymbol{A}_\downarrow := \boldsymbol{I}_\downarrow\boldsymbol{A}_L^\top$. Then, we observe $\boldsymbol{U}_\uparrow = \boldsymbol{A}_\uparrow\boldsymbol{P}$, $\boldsymbol{U}_\downarrow = \boldsymbol{A}_\downarrow\boldsymbol{P}$ and particularly for the Vandermonde matrix $\boldsymbol{A}_\downarrow = \boldsymbol{A}_\uparrow \operatorname{diag}(z_1, \ldots, z_M)$. This leads to

$$\boldsymbol{U}_\downarrow = \boldsymbol{A}_\downarrow\boldsymbol{P} = \boldsymbol{A}_\uparrow \operatorname{diag}(z_1, \ldots, z_M)\boldsymbol{P} = \boldsymbol{U}_\uparrow\boldsymbol{P}^{-1} \operatorname{diag}(z_1, \ldots, z_M)\boldsymbol{P}.$$

Therefore, we can define $\boldsymbol{X} := \boldsymbol{P}^{-1} \operatorname{diag}(z_1, \ldots, z_M)\boldsymbol{P}$ and we directly see, by similarity transformation, that the eigenvalues of $\boldsymbol{X}$ are equal to the nodes $\Lambda = \{z_1, \ldots, z_M\}$. Furthermore, since $L - 1 \geq M$, we have $\operatorname{rank}(\boldsymbol{U}_\uparrow) = \operatorname{rank}(\boldsymbol{A}_{L-1}) = M$ and $\boldsymbol{X}$ is uniquely given by

$$\boldsymbol{X} = \boldsymbol{U}_\uparrow^\dagger \boldsymbol{U}_\downarrow. \tag{4.2.6}$$

At this point only matrices are involved, from which we can calculate the given data. In a second step the coefficients $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_M)^\top$ of the exponential sum (4.1.1) can be reconstructed from the sample vector $\boldsymbol{\mathcal{E}} = (\mathcal{E}_0, \ldots, \mathcal{E}_{N-1})$ with the calculation

$$\boldsymbol{\alpha} = \left(\boldsymbol{A}_N^T\right)^\dagger \boldsymbol{\mathcal{E}}. \tag{4.2.7}$$

In practice, one usually deals with the perturbed problem (4.1.2). Therefore, we get a perturbed Hankel matrix

$$\widetilde{\boldsymbol{H}}_L := \boldsymbol{H}_L + \boldsymbol{E} := \left(\widetilde{\mathcal{E}}_{k+\ell}\right)_{k=0,l=0}^{L-1,N-L}.$$

To make sure that this matrix has at least rank M, the constraint

$$\|\boldsymbol{E}\| < \sigma_M(\boldsymbol{H}_L)$$

is imposed. Together with Lemma 2.1.16, this yields $\sigma_M(\widetilde{\boldsymbol{H}}_L) \geq \sigma_M(\boldsymbol{H}) - \|\boldsymbol{E}\| = \sigma_M(\boldsymbol{H}) - \|\boldsymbol{E}\| > 0$. However, the perturbed matrix $\boldsymbol{H}_L$ will generically have full rank, so that we use its M-truncated SVD $\widetilde{\boldsymbol{U}}_0\widetilde{\boldsymbol{\Sigma}}_M\widetilde{\boldsymbol{V}}_0^*$ instead. That means for the ESPRIT algorithm, we take the matrix $\widetilde{\boldsymbol{U}}$ into account which has the first M singular vectors (first M columns of $\widetilde{\boldsymbol{U}}_0$) as columns. From there on, we proceed with the ESPRIT algorithm as in the unperturbed case. Note that calculations where the Moore-Penrose pseudo inverse is involved become least squares solutions to the respective over-determined linear system. In the following, the involved matrices for the perturbed case are denoted as before with an additional "$\widetilde{\ }$" on top. We end up with a matrix $\widetilde{\boldsymbol{X}} \in \mathbb{C}^{M\times M}$ that has the reconstructed nodes $\widetilde{\Lambda} = \{\widetilde{z}_1, \ldots, \widetilde{z}_M\}$ as eigenvalues. These nodes do not lay on the unit circle necessarily. In order to reconstruct the corresponding nodes $\widetilde{\Omega} = \left\{\widetilde{t}_1, \ldots, \widetilde{t}_M\right\} \subset \mathbb{T}$, we simply take the respective normalized complex arguments

$$\widetilde{t}_j := \frac{\arg_{[0,2\pi)}(z_j)}{2\pi},$$

where

$$\arg_{[0,2\pi)} \colon \mathbb{C} \to [0, 2\pi), \quad \arg_{[0,2\pi)}(z) := \begin{cases} \arccos\left(\frac{\mathrm{Re}(z)}{|z|}\right), & \mathrm{Im}(z) \geq 0, \\ 2\pi - \arccos\left(\frac{\mathrm{Re}(z)}{|z|}\right), & \mathrm{Im}(z) < 0. \end{cases}$$

This can be seen as a projection onto the circle. The ESPRIT algorithm for perturbed data is summed up in Algorithm 1.

In the next section, we analyze how close the reconstructed nodes from the perturbed problem are to the original ones depending on the perturbation $\boldsymbol{E}$.

4.3 Stability of the node reconstruction

In this section, we present the stability analysis of the node reconstruction by the ESPRIT algorithm. More precisely, we want to understand, in how far reconstructed nodes differ from the original ones when the given data is corrupted. We orient mostly at [74] and provide a comparison to results from [5, 6] in the end.

Algorithm 1: ESPRIT

Input: perturbed samples $\widetilde{\boldsymbol{\mathcal{E}}} = (\widetilde{\mathcal{E}}_0, \dots, \widetilde{\mathcal{E}}_{N-1})^\top \in \mathbb{C}^N$ and order M of an exponential sum
Output: nodes $\widetilde{\Omega} = \{\widetilde{t}_1, \dots, \widetilde{t}_M\}$ and coefficients $\widetilde{\alpha}_1, \dots, \widetilde{\alpha}_M$
begin
Step 1 - reconstruction of nodes
choose $M < L \leq N - M + 1$
set up the Hankel matrix $\widetilde{\boldsymbol{H}}_L = \left(\widetilde{\mathcal{E}}_{k+\ell}\right)_{k,\ell=0}^{L-1,N-L}$
compute the economic version of the M-truncated SVD $\widetilde{\boldsymbol{H}}_L = \widetilde{\boldsymbol{U}}\widetilde{\boldsymbol{\Sigma}}\widetilde{\boldsymbol{V}}$
compute $\widetilde{\boldsymbol{X}} = \widetilde{\boldsymbol{U}}_\uparrow^\dagger \widetilde{\boldsymbol{U}}_\downarrow$, where $\widetilde{\boldsymbol{U}}_\uparrow, \widetilde{\boldsymbol{U}}_\downarrow$ have the respective first and last $L-1$ rows of $\boldsymbol{U}$
compute the eigenvalues $\widetilde{z}_j, j = 1, \dots, M$ of $\widetilde{\boldsymbol{X}}$
for $j = 1, \dots, M$ **do**
  compute the nodes $\widetilde{t}_j = \arg_{[0,2\pi)}(\widetilde{z}_j)$
Step 2 - reconstruction of coefficients
set up the Vandermonde matrix $\widetilde{\boldsymbol{A}}_N = \left(\mathrm{e}^{2\pi i \widetilde{t}_j (k-1)}\right)_{j,k=1}^{M,N}$
compute the coefficient vector $\widetilde{\boldsymbol{\alpha}} = \left(\widetilde{\boldsymbol{A}}_N^\top\right)^\dagger \widetilde{\boldsymbol{\mathcal{E}}}$

4.3.1 Eigenvalue perturbation

Lemma 4.3.1 (Connection between matching distances, cf. [74, App. B])**.**
For $\Omega = \{t_1, \dots, t_M\}, \widetilde{\Omega} = \{\widetilde{t}_1, \dots, \widetilde{t}_M\} \subset \mathbb{T}$ define the matching distance *analogous to Definition 2.2.1 by*

$$\mathrm{md}_{\mathbb{T}}(\Omega, \widetilde{\Omega}) := \min_{\pi} \max_{1 \leq j \leq M} \left| t_j - t_{\pi(j)} \right|_{\mathbb{T}},$$

where $\pi\colon \{1, \dots, M\} \to \{1, \dots, M\}$ is a permutation. Let $\Omega, \widetilde{\Omega} \subset \mathbb{T}$ and $\Lambda, \widetilde{\Lambda} \subset \mathbb{C}$ given as in Section 4.2, i.e. Λ and $\widetilde{\Lambda}$ contain complex numbers where the ones of Λ lie on the unit circle. Ω and $\widetilde{\Omega}$ are the corresponding parameter in $\mathbb{T}$. Then we have

$$\mathrm{md}_{\mathbb{T}}(\Omega, \widetilde{\Omega}) \leq \frac{1}{2} \mathrm{md}(\Lambda, \widetilde{\Lambda}).$$

Proof. The proof can be found in the given reference and is done by geometric arguments and use of the sine law. We add here, that the constant can not be improved, by looking at the following example. Let $M = 1$ and $z = 1$, $\widetilde{z}_\epsilon = -\epsilon$ for some $0 \leq \epsilon < 1$. Then we have

$$\mathrm{md}_{\mathbb{T}}(\Omega, \widetilde{\Omega}) = \left| t - \widetilde{t}_\epsilon \right|_{\mathbb{T}} = \frac{1}{2} \leq \frac{1}{2}(1 + \epsilon) = \frac{1}{2} \left| z - \widetilde{z}_\epsilon \right| = \mathrm{md}(\Lambda, \widetilde{\Lambda})$$

for all $\epsilon > 0$. Thus, a constant smaller than $\frac{1}{2}$ is not possible. □

We apply Wedin's theorem for our particular case similar to [74, La. 5] followed by an additional lemma for preparing the next theorem.

Lemma 4.3.2 (Wedin and principal vector matrices)**.**
Let the notation and the data be as in the ESPRIT algorithm, Section 4.2. Then there exist unitary matrices $\boldsymbol{R}, \boldsymbol{S} \in \mathbb{C}^{M \times M}$ such that, if $\|\boldsymbol{E}\| \leq \frac{\sigma_M(\boldsymbol{H}_L)}{2}$, we have

$$\left\| \boldsymbol{U}\boldsymbol{R} - \widetilde{\boldsymbol{U}}\boldsymbol{S} \right\| \leq \frac{2\sqrt{2}\,\|\boldsymbol{E}\|}{\sigma_M(\boldsymbol{H}_L)}.$$

Proof. Firstly, the assumption imply $\|\boldsymbol{E}\| < \sigma_M(\boldsymbol{H}_L)$ and therefore, $\widetilde{\sigma}_M(\widetilde{\boldsymbol{H}}_L) \geq \sigma_M(\boldsymbol{H}_L) - \|\boldsymbol{E}\| > 0$ by Lemma 2.1.16. This guarantees that $\operatorname{rank}(\boldsymbol{H}_L) \geq M$ and thus, $\dim \operatorname{range}(\widetilde{\boldsymbol{U}}) = \operatorname{rank}(\widetilde{\boldsymbol{U}}) = M$. Furthermore, the columns of $\boldsymbol{U}$ and $\widetilde{\boldsymbol{U}}$ are orthonormal bases for their respective ranges by construction. Let $0 \leq \theta_1 \leq \cdots \leq \theta_M \leq \frac{\pi}{2}$ be the principal angles and $\boldsymbol{w}_j, \widetilde{\boldsymbol{w}}_j, j = 1, \ldots, M$ be the principal vectors between $\operatorname{range}(\boldsymbol{U})$ and $\operatorname{range}(\widetilde{\boldsymbol{U}})$ given by Definition 2.2.4. According to Lemma 2.1.19, we find unitary change of bases matrices $\boldsymbol{R}, \boldsymbol{S} \in \mathbb{C}^{M\times M}$, such that $\boldsymbol{W} := [\boldsymbol{w}_1, \ldots, \boldsymbol{w}_M] = \boldsymbol{U}\boldsymbol{R}$ and $\widetilde{\boldsymbol{W}} := [\widetilde{\boldsymbol{w}}_1, \cdots, \widetilde{\boldsymbol{w}}_M] = \widetilde{\boldsymbol{U}}\boldsymbol{S}$. Now, Lemma 2.2.8 provides

$$\left\|\boldsymbol{W} - \widetilde{\boldsymbol{W}}\right\| \leq \sqrt{2}\sin(\theta_M). \tag{4.3.1}$$

This bound is in terms of the largest principal angle and since the matrices $\boldsymbol{U}, \widetilde{\boldsymbol{U}}$ come from the SVD of $\boldsymbol{H}_L$ and $\widetilde{\boldsymbol{H}}_L$, Wedin's Theorem (Theorem 2.2.9) helps. We apply it to the matrices $\boldsymbol{H}_L$ and $\widetilde{\boldsymbol{H}}_L$ with the subspaces $\operatorname{range}(\boldsymbol{U}), \operatorname{range}(\widetilde{\boldsymbol{U}})$ and the parameter $k = M$. The assumption $\|\boldsymbol{E}\| \leq \frac{\sigma_M(\boldsymbol{H}_L)}{2}$ allows to set $\gamma = \sigma_M(\boldsymbol{H}_L) - \|\boldsymbol{E}\| = \frac{1}{2}\sigma_M(\boldsymbol{H}_L) > 0$. Furthermore, Lemma 2.1.16 guarantees with $\sigma_{M+1}(\boldsymbol{H}_L) = 0$, that $\sigma_{M+1}(\widetilde{\boldsymbol{H}}_L) \leq \|\boldsymbol{E}\|$ and together with the assumptions on $\|\boldsymbol{E}\|$ also $\operatorname{rank}(\widetilde{\boldsymbol{H}}_L) \geq M$. Consequently, the choice of the parameters $\beta = \|\boldsymbol{E}\|$ and $k = M$ is valid. We have clearly $\sigma_M(\boldsymbol{H}_L) = \gamma + \beta$ which is the last assumption of Wedin's Theorem. Additionally, we observe $\|\boldsymbol{V}\| = \|\boldsymbol{U}^*\| = 1$, and hence, $\|\boldsymbol{E}\boldsymbol{V}\| \leq \|\boldsymbol{E}\|\,\|\boldsymbol{V}\| = \|\boldsymbol{E}\|$ and $\|\boldsymbol{U}^*\boldsymbol{E}\| \leq \|\boldsymbol{E}\|$, analogously. Then the application of Wedin's theorem yields

$$\sin(\theta_M) = \|\sin(\boldsymbol{\theta})\| \leq \frac{\max\{\|\boldsymbol{E}\boldsymbol{V}\|, \|\boldsymbol{U}^*\boldsymbol{E}\|\}}{\sigma_M(\boldsymbol{H}_L) - \|\boldsymbol{E}\|} \leq \frac{\|\boldsymbol{E}\|}{\sigma_M(\boldsymbol{H}_L) - \|\boldsymbol{E}\|} \leq \frac{2\,\|\boldsymbol{E}\|}{\sigma_M(\boldsymbol{H}_L)}$$

and together with (4.3.1) the result. □

Lemma 4.3.3.
Let the notation and the data be as in the ESPRIT algorithm, Section 4.2, and let $\boldsymbol{R}, \boldsymbol{S} \in \mathbb{C}^{M\times M}$ be the matrices given by Lemma 4.3.2. If $\left\|\boldsymbol{U}\boldsymbol{R} - \widetilde{\boldsymbol{U}}\boldsymbol{S}\right\| \leq \frac{\sigma_{\min}(\boldsymbol{U}_\uparrow)}{2}$, then

$$\left\|(\boldsymbol{U}_\uparrow\boldsymbol{R})^\dagger - (\widetilde{\boldsymbol{U}}_\uparrow\boldsymbol{S})^\dagger\right\| \leq \frac{2\sqrt{2}\left\|\boldsymbol{U}\boldsymbol{R} - \widetilde{\boldsymbol{U}}\boldsymbol{S}\right\|}{\sigma_{\min}(\boldsymbol{U}_\uparrow)^2}.$$

Proof. We apply Theorem 2.2.11 after verifying the assumptions. Since the matrices $\boldsymbol{R}$ and $\boldsymbol{P}$ are regular, $\operatorname{rank}(\boldsymbol{U}_\uparrow\boldsymbol{R}) = \operatorname{rank}(\boldsymbol{U}_\uparrow) = \operatorname{rank}(\boldsymbol{A}_\uparrow\boldsymbol{P}) = \operatorname{rank}(\boldsymbol{A}_\uparrow) = \operatorname{rank}(\boldsymbol{A}_{L-1}^\top) = M$ and thus, the matrix $\boldsymbol{U}_\uparrow\boldsymbol{R}$ has full rank. For any matrix $\boldsymbol{B} \in \mathbb{C}^{L\times M}$ and $\boldsymbol{B}_\uparrow := \boldsymbol{I}_\uparrow\boldsymbol{B} \in \mathbb{C}^{(L-1)\times M}$, with $\boldsymbol{I}_\uparrow$ given by (4.2.5), we get, by Lemma 2.1.18

$$\|\boldsymbol{B}_\uparrow\| \leq \|\boldsymbol{B}\|. \tag{4.3.2}$$

Applied to $\boldsymbol{U}_\uparrow\boldsymbol{R} - \widetilde{\boldsymbol{U}}_\uparrow\boldsymbol{S} = (\boldsymbol{U}\boldsymbol{R} - \widetilde{\boldsymbol{U}}\boldsymbol{S})_\uparrow$ and using the assumptions leads to

$$\left\|\boldsymbol{U}_\uparrow\boldsymbol{R} - \widetilde{\boldsymbol{U}}_\uparrow\boldsymbol{S}\right\| \leq \left\|\boldsymbol{U}\boldsymbol{R} - \widetilde{\boldsymbol{U}}\boldsymbol{S}\right\| \leq \frac{\sigma_{\min}(\boldsymbol{U}_\uparrow)}{2} = \frac{\sigma_M(\boldsymbol{U}_\uparrow)}{2}. \tag{4.3.3}$$

From this it follows by Lemma 2.1.16, $\boldsymbol{R}$ being unitary, and $\boldsymbol{U}_\uparrow$ having full rank (equivalent to $\sigma_M(\boldsymbol{U}_\uparrow) > 0$)

$$\sigma_M(\widetilde{\boldsymbol{U}}_\uparrow\boldsymbol{S}) \geq \sigma_M(\boldsymbol{U}_\uparrow\boldsymbol{R}) - \left\|\boldsymbol{U}_\uparrow\boldsymbol{R} - \widetilde{\boldsymbol{U}}_\uparrow\boldsymbol{S}\right\| \geq \sigma_M(\boldsymbol{U}_\uparrow) - \frac{1}{2}\sigma_M(\boldsymbol{U}_\uparrow) = \frac{1}{2}\sigma_M(\boldsymbol{U}_\uparrow) > 0. \tag{4.3.4}$$

Hence, $\widetilde{\boldsymbol{U}}_{\uparrow}\boldsymbol{S}$ has full rank M and the assumptions of Theorem 2.2.11 are satisfied. Then its application together with (4.3.4) and $\boldsymbol{R}, \boldsymbol{S}$ being unitary yields

$$\left\|(\boldsymbol{U}_{\uparrow}\boldsymbol{R})^{\dagger} - (\widetilde{\boldsymbol{U}}_{\uparrow}\boldsymbol{S})^{\dagger}\right\| \leq \frac{\sqrt{2}\left\|\boldsymbol{U}_{\uparrow}\boldsymbol{R} - \widetilde{\boldsymbol{U}}_{\uparrow}\boldsymbol{S}\right\|}{\sigma_{\min}(\boldsymbol{U}_{\uparrow}\boldsymbol{R})\,\sigma_{\min}(\widetilde{\boldsymbol{U}}_{\uparrow}\boldsymbol{S})} \leq \frac{2\sqrt{2}\left\|\boldsymbol{U}_{\uparrow}\boldsymbol{R} - \widetilde{\boldsymbol{U}}_{\uparrow}\boldsymbol{S}\right\|}{\sigma_{\min}(\boldsymbol{U}_{\uparrow})^2}.$$

□

Remark 4.3.4.
In [74], the authors used the special case of Theorem 3.2 in [52] to bound $\left\|(\boldsymbol{U}_{\uparrow}\boldsymbol{R})^{\dagger} - (\widetilde{\boldsymbol{U}}_{\uparrow}\boldsymbol{S})^{\dagger}\right\|$. *However, the proof of that special case in [52] seems to be imprecise and inconsistent with the rest of the proof. To make it clear, if* $\boldsymbol{M} \in \mathbb{C}^{m\times n}$ *and* $\widetilde{\boldsymbol{M}} \in \mathbb{C}^{m\times n}$ *denotes its perturbed version, then [52, Thm. 3.2] provides an upper bound for* $\left\|(\boldsymbol{M}_k)^{\dagger} - (\widetilde{\boldsymbol{M}}_k)^{\dagger}\right\|$, *where* $\boldsymbol{M}_k$ *and* $\widetilde{\boldsymbol{M}}_k$ *are the corresponding* k*-truncated SVD matrices. The special case is for* $k = r = \operatorname{rank}(\boldsymbol{M})$, *i.e. we have* $\boldsymbol{M}_k = \boldsymbol{M}_r$, $\sigma_r(\boldsymbol{M}_r) > 0$, $\sigma_{r+1}(\boldsymbol{M}_r) = 0$ *and since* $\left\|\boldsymbol{M} - \widetilde{\boldsymbol{M}}\right\| < \sigma_r(\boldsymbol{M})$ *by assumption, Lemma 2.1.16 provides* $\operatorname{rank}(\widetilde{\boldsymbol{M}}) \geq r$. *In [52, Thm. 3.2] the deduction,* $\operatorname{rank}(\widetilde{\boldsymbol{M}}) = \operatorname{rank}(\boldsymbol{M})(= r)$ *and therefore* $\widetilde{\sigma}_{r+1} = 0$, *is made. Confusion comes now from the unfortunate situation that it is not clearly stated whether* $\widetilde{\sigma}_j$ *denotes the* j*-th singular value of* $\widetilde{\boldsymbol{M}}$ *or* $\widetilde{\boldsymbol{M}}_k$ *for some index* j. *If* $\widetilde{\sigma}_j = \widetilde{\sigma}_j(\widetilde{\boldsymbol{M}}_k)$ *then the above deduction is trivial, because the* $r+1$*-st singular value is, by construction of the* r*-truncated SVD matrix, zero. If* $\widetilde{\sigma}_j = \widetilde{\sigma}_j(\widetilde{\boldsymbol{M}})$ *the deduction is only true for the case* $\boldsymbol{M}$ *has full rank, i.e.* $r = \min\{m, n\}$. *We assume the latter is meant since in the remaining proof of that theorem* $\widetilde{\sigma}_{k+1}$ *is treated to be nonzero.*

To conclude, the statement of the special case is only true, if $\boldsymbol{M}$ *is a full rank matrix, otherwise doing the necessary modifications in the proof would lead to the statement*

$$\frac{\left\|\boldsymbol{M}^{\dagger} - \widetilde{\boldsymbol{M}}_k^{\dagger}\right\|}{\left\|\boldsymbol{M}^{\dagger}\right\|} \leq \frac{3\sigma_1(\boldsymbol{M})}{\left(1 - \frac{\|\boldsymbol{M}-\widetilde{\boldsymbol{M}}\|}{\sigma_r(\boldsymbol{M})}\right)^2 \sigma_r(\boldsymbol{M})} \cdot \frac{\left\|\boldsymbol{M} - \widetilde{\boldsymbol{M}}\right\|}{\|\boldsymbol{M}\|}.$$

Here, we avoid these problems by applying Theorem 2.2.11 instead. We obtain the constant $\sqrt{2}$ *as benefit, which is about a factor* 2 *better than the constant* 3 *in [74, p. 13].*

Theorem 4.3.5 (Error amplification for nodes)**.**
Let the notation and the data be as in the ESPRIT algorithm, Section 4.2 and define $\alpha_{\min} := \min_{j=1,\cdots,M} |\alpha_j|$. *If the perturbation* $\boldsymbol{E}$ *on the Hankel matrix* $\boldsymbol{H}_L$ *satisfies*

$$\|\boldsymbol{E}\| \leq \frac{\alpha_{\min}\,\sigma_{\min}(\boldsymbol{A}_L)\,\sigma_{\min}(\boldsymbol{A}_{N-L+1})\,\sigma_{\min}(\boldsymbol{U}_{\uparrow})}{4\sqrt{2}},$$

then the matching distance (Definition 2.2.1) between the nodes Ω *and the reconstructed nodes* $\widetilde{\Omega}$ *from the ESPRIT algorithm is bounded by*

$$\operatorname{md}_{\mathbb{T}}(\Omega, \widetilde{\Omega}) \leq \frac{6(2M-1)\,\|\boldsymbol{A}_L\|\,\|\boldsymbol{E}\|}{\alpha_{\min}\,\sigma_{\min}(\boldsymbol{A}_L)^2\,\sigma_{\min}(\boldsymbol{A}_{N-L+1})\,\sigma_{\min}(\boldsymbol{U}_{\uparrow})^2}.$$

Proof. We apply Theorem 2.2.3. But instead of applying it directly to the matrix $\boldsymbol{X} = \boldsymbol{P}^{-1}\operatorname{diag}(z_1, \ldots, z_M)\boldsymbol{P}$, which is diagonalizable per construction, and its perturbed version

$\widetilde{\boldsymbol{X}}$, we apply it to similarity transformed versions. The reason is, that directly applying would lead to the task to upper bound $\|\boldsymbol{U} - \widetilde{\boldsymbol{U}}\|$. This turns out to be difficult since in general only few information is known about $\boldsymbol{U}$ and $\widetilde{\boldsymbol{U}}$ except for having orthonormal columns. Therefore, we use the fact, that eigenvalues are invariant under similarity transformations. Let $\boldsymbol{R}, \boldsymbol{S} \in \mathbb{C}^{M\times M}$ be the unitary matrices given by Lemma 4.3.2. Then $\boldsymbol{R}^{-1}\boldsymbol{X}\boldsymbol{R}$ and $\boldsymbol{S}^{-1}\widetilde{\boldsymbol{X}}\boldsymbol{S}$ still have the same eigenvalues as $\boldsymbol{X}$ and $\widetilde{\boldsymbol{X}}$. Hence, Theorem 2.2.3 together with Lemma 4.3.1 yields

$$\mathrm{md}_{\mathbb{T}}(\Omega, \widetilde{\Omega}) \leq \frac{1}{2}(2M-1)\frac{\sigma_{\max}(\boldsymbol{R}\boldsymbol{P})}{\sigma_{\min}(\boldsymbol{R}\boldsymbol{P})} \left\| \boldsymbol{R}^{-1}\boldsymbol{X}\boldsymbol{R} - \boldsymbol{S}^{-1}\widetilde{\boldsymbol{X}}\boldsymbol{S} \right\|. \tag{4.3.5}$$

Since $\boldsymbol{R}$ is unitary all its singular values equal one. Furthermore, $\boldsymbol{A}_L$ has full rank so that $\boldsymbol{P} = (\boldsymbol{A}_L^{\top})^{\dagger}\boldsymbol{U}$. Lemma 2.1.19 gives $\sigma_{\min}(\boldsymbol{U}) \leq \sigma_{\max}(\boldsymbol{U}) = 1$ and therefore, applying Lemma 2.1.16 yields

$$\sigma_{\max}(\boldsymbol{R}\boldsymbol{P}) \leq \|\boldsymbol{R}\|\,\|\boldsymbol{P}\| \leq \left\|(\boldsymbol{A}_L^{\top})^{\dagger}\right\|\|\boldsymbol{U}\| = \frac{1}{\sigma_{\min}(\boldsymbol{A}_L)}, \tag{4.3.6}$$

and

$$\sigma_{\min}(\boldsymbol{R}\boldsymbol{P}) \geq \sigma_{\min}(\boldsymbol{P}) \geq \sigma_{\min}(\boldsymbol{A}_L^{\dagger})\,\sigma_{\min}(\boldsymbol{U}) = \frac{1}{\sigma_{\max}(\boldsymbol{A}_L)}. \tag{4.3.7}$$

Now, we bound $\left\|\boldsymbol{R}^{-1}\boldsymbol{X}\boldsymbol{R} - \boldsymbol{S}^{-1}\widetilde{\boldsymbol{X}}\boldsymbol{S}\right\|$. Analogously to (4.3.2) and (4.3.3), we have

$$\left\|\boldsymbol{U}_{\downarrow}\boldsymbol{R} - \widetilde{\boldsymbol{U}}_{\downarrow}\boldsymbol{S}\right\| \leq \left\|\boldsymbol{U}\boldsymbol{R} - \widetilde{\boldsymbol{U}}\boldsymbol{S}\right\| \quad \text{and} \quad \left\|\widetilde{\boldsymbol{U}}_{\downarrow}\boldsymbol{S}\right\| \leq \left\|\widetilde{\boldsymbol{U}}\right\| = 1. \tag{4.3.8}$$

The inequality $\sigma_{\min}(\boldsymbol{U}_{\uparrow}) \leq \sigma_{\max}(\boldsymbol{U}_{\uparrow}) = 1$ holds and Lemma 2.1.16 together with (4.2.3) imply

$$\sigma_M(\boldsymbol{H}_L) \geq \alpha_{\min}\,\sigma_{\min}(\boldsymbol{A}_L)\,\sigma_{\min}(\boldsymbol{A}_{N-L+1}). \tag{4.3.9}$$

Thus, the assumption on $\|\boldsymbol{E}\|$ ensures that $\|\boldsymbol{E}\| \leq \frac{1}{2}\sigma_M(\boldsymbol{H}_L)$. Therefore, we can apply Lemma 4.3.2 and reuse the assumption on $\|\boldsymbol{E}\|$ to get

$$\left\|\boldsymbol{U}\boldsymbol{R} - \widetilde{\boldsymbol{U}}\boldsymbol{S}\right\| \leq \frac{2\sqrt{2}\,\|\boldsymbol{E}\|}{\sigma_M(\boldsymbol{H}_L)} \leq \frac{\sigma_{\min}(\boldsymbol{U}_{\uparrow})}{2}. \tag{4.3.10}$$

This, in turn, is required in Lemma 4.3.3, which we now can apply to obtain

$$\begin{aligned}
\left\|\boldsymbol{R}^{-1}\boldsymbol{X}\boldsymbol{R} - \boldsymbol{S}^{-1}\widetilde{\boldsymbol{X}}\boldsymbol{S}\right\| &\leq \left\|\boldsymbol{R}^{-1}\boldsymbol{U}_{\uparrow}^{\dagger}\boldsymbol{U}_{\downarrow}\boldsymbol{R} - \boldsymbol{R}^{-1}\boldsymbol{U}_{\uparrow}^{\dagger}\widetilde{\boldsymbol{U}}_{\downarrow}\boldsymbol{S}\right\| + \left\|\boldsymbol{R}^{-1}\boldsymbol{U}_{\uparrow}^{\dagger}\widetilde{\boldsymbol{U}}_{\downarrow}\boldsymbol{S} - \boldsymbol{S}^{-1}\widetilde{\boldsymbol{U}}_{\uparrow}^{\dagger}\widetilde{\boldsymbol{U}}_{\downarrow}\boldsymbol{S}\right\| \\
&\leq \left\|\boldsymbol{R}^{-1}\boldsymbol{U}_{\uparrow}^{\dagger}\right\|\left\|\boldsymbol{U}_{\downarrow}\boldsymbol{R} - \widetilde{\boldsymbol{U}}_{\downarrow}\boldsymbol{S}\right\| + \left\|\boldsymbol{R}^{-1}\boldsymbol{U}_{\uparrow}^{\dagger} - \boldsymbol{S}^{-1}\widetilde{\boldsymbol{U}}_{\uparrow}^{\dagger}\right\|\left\|\widetilde{\boldsymbol{U}}_{\downarrow}\boldsymbol{S}\right\| \\
&\leq \frac{\left\|\boldsymbol{U}_{\downarrow}\boldsymbol{R} - \widetilde{\boldsymbol{U}}_{\downarrow}\boldsymbol{S}\right\|}{\sigma_{\min}(\boldsymbol{U}_{\uparrow})} + \frac{2\sqrt{2}\left\|\boldsymbol{U}\boldsymbol{R} - \widetilde{\boldsymbol{U}}\boldsymbol{S}\right\|}{\sigma_{\min}(\boldsymbol{U}_{\uparrow})^2} \leq \frac{(1+2\sqrt{2})\left\|\boldsymbol{U}\boldsymbol{R} - \widetilde{\boldsymbol{U}}\boldsymbol{S}\right\|}{\sigma_{\min}(\boldsymbol{U}_{\uparrow})^2}.
\end{aligned}$$

After applying the first inequality of (4.3.10) and (4.3.9), putting everything together in (4.3.5) yields the result. □

Remark 4.3.6 (Comparison to [6, 5])**.**
In Lemmata 4.3.2 and 4.3.3 the assumptions on the error matrix $\boldsymbol{E}$ include a factor of $\frac{1}{2}$, whereas a strict inequality without the factor would have been sufficient. This means we payed

a slightly worse constant in order to get results that are structured more clearly by avoiding terms of the form $\sigma_M(\boldsymbol{H}_L) - \|\boldsymbol{E}\|$ in Lemma 4.3.2.

This was not done in [5], so that the result stated there seems to be hardly comparable to ours. But when using the same principle in [5] and additionally stopping the analysis in there without bounding $\sigma_{\min}(\boldsymbol{U}_\uparrow)$, we obtain the following. The details are in Appendix C. If the error matrix fulfills

$$\|\boldsymbol{E}\|_{\mathrm{F}} \leq \frac{\alpha_{\min}\, \sigma_{\min}(\boldsymbol{A}_L)\, \sigma_{\min}(\boldsymbol{A}_{N-L+1})\, \sigma_{\min}(\boldsymbol{U}_\uparrow)}{4\sqrt{2}},$$

then we have

$$\mathrm{md}_{\mathbb{T}}(\Omega, \widetilde{\Omega}) \leq \frac{6(2M-1)\, \|\boldsymbol{A}_L\|\, \|\boldsymbol{E}\|_{\mathrm{F}}}{\alpha_{\min}\, \sigma_{\min}(\boldsymbol{A}_L)^2\, \sigma_{\min}(\boldsymbol{A}_{N-L+1})\, \sigma_{\min}(\boldsymbol{U}_\uparrow)^2}.$$

which is due to the Frobenius norm only slightly weaker than our result in Theorem 4.3.5. This shows that [74], which we followed, provides the same result as [5] in principle (after the improvement done here with respect to the dependency on the number of nodes).

Remark 4.3.7 (Worst case error analysis)**.**
The analysis presented here is of worst case nature and all nodes are treated equally. Recent development in the literature [13] suggest that in case of clustered node configurations the reconstruction of isolated nodes is way less difficult than reconstruction of nodes inside clusters. Additionally, [12] analyzed the whole set of singular values of Vandermonde matrices with clustered node configurations. The authors showed that for each cluster there is a singular value that scales like $(qN)^{\ell-1}$, where ℓ is the number of nodes in that cluster, see Theorems 3.4.24 and 3.4.25. Only singular values corresponding to clusters with biggest cardinality behave as bad as the bounds for the smallest singular value of the Vandermonde suggest. So stability analysis of the ESPRIT algorithm for the reconstruction of specific nodes would be of high interest and could benefit from these results.

4.3.2 Lower bounds on $\sigma_{\min}(\boldsymbol{U}_\uparrow)$

We saw that the smallest singular value of $\boldsymbol{U}_\uparrow$ plays an important role for the performance analysis of the ESPRIT algorithm. Since $\boldsymbol{U}_\uparrow$ is constructed by deleting the last row of the matrix $\boldsymbol{U}$, which comes from SVD of the Hankel matrix and has orthonormal columns, we directly know by Lemma 2.1.18 that $0 \leq \sigma_{\min}(\boldsymbol{U}_\uparrow) \leq 1$, cf. (4.3.2). The following lemma allows a more precise description.

Lemma 4.3.8 (cf. [74, App. D])**.**
Let $L, M \in \mathbb{N}_+$, $L > M$. We define

$$\gamma := \sup_{\mathrm{supp}(\mu) \subset \Omega} \frac{|\widehat{\mu}(0)|^2}{\sum_{k=0}^{L-1} |\widehat{\mu}(k)|^2},$$

where $\mu = \sum_{j=1}^{M} c_j \delta_{t_j}$ is a discrete measure (see Definition 2.3.5), with some coefficients $c_1, \ldots, c_M \in \mathbb{C}$ and $t_j \in \Omega \subset \mathbb{T}$ being the nodes corresponding to the Vandermonde matrix $\boldsymbol{A}_L$. Then a lower bound for $\sigma_{\min}(\boldsymbol{U}_\uparrow)$ is given by

$$\sigma_{\min}(\boldsymbol{U}_\uparrow) \geq \sqrt{1-\gamma}.$$

Proof. We can write $\boldsymbol{U} = \begin{pmatrix} \boldsymbol{U}_\uparrow \\ \boldsymbol{w}^* \end{pmatrix}$ with the vector $\boldsymbol{w} \in \mathbb{C}^M$. Since $\boldsymbol{U}$ has orthogonal columns, we have

$$\boldsymbol{I}_M = \boldsymbol{U}^*\boldsymbol{U} = \begin{pmatrix} \boldsymbol{U}_\uparrow^* & \boldsymbol{w} \end{pmatrix} \begin{pmatrix} \boldsymbol{U}_\uparrow \\ \boldsymbol{w}^* \end{pmatrix} = \boldsymbol{U}_\uparrow^*\boldsymbol{U}_\uparrow + \boldsymbol{w}\boldsymbol{w}^*$$

and therefore,

$$\boldsymbol{U}^*\boldsymbol{U} = \boldsymbol{I}_M - \boldsymbol{w}\boldsymbol{w}^*.$$

Applying this matrix to the vector $\boldsymbol{w}$ yields the eigenvalue equation

$$\boldsymbol{U}_\uparrow^*\boldsymbol{U}_\uparrow\boldsymbol{w} = \boldsymbol{I}_M\boldsymbol{w} + \boldsymbol{w}\boldsymbol{w}^*\boldsymbol{w} = \boldsymbol{w}(1 - \|\boldsymbol{w}\|^2).$$

Thus, $\boldsymbol{w}$ is an eigenvector of $\boldsymbol{U}_\uparrow^*\boldsymbol{U}_\uparrow$ with eigenvalue $1 - \|\boldsymbol{w}\|^2$. Furthermore, each vector $\boldsymbol{w}_\perp$ perpendicular to $\boldsymbol{w}$ is, by

$$\boldsymbol{U}_\uparrow^*\boldsymbol{U}_\uparrow\boldsymbol{w}_\perp = \boldsymbol{I}_M\boldsymbol{w}_\perp + \boldsymbol{w}\boldsymbol{w}^*\boldsymbol{w}_\perp = \boldsymbol{w}_\perp,$$

eigenvector of $\boldsymbol{U}_\uparrow^*\boldsymbol{U}_\uparrow$ with eigenvalue 1. It follows, that the singular values of $\boldsymbol{U}_\uparrow$ are given by

$$\sigma_1(\boldsymbol{U}_\uparrow) = \cdots = \sigma_{M-1}(\boldsymbol{U}_\uparrow) = 1 \text{ and } \sigma_{\min}(\boldsymbol{U}_\uparrow) = \sqrt{1 - \|\boldsymbol{w}\|^2}.$$

Since singular values are non-negative we have $\|\boldsymbol{w}\| \leq 1$. Hence, the task is to provide an upper bound for $\|\boldsymbol{w}\|$ which is better than the trivial bound 1.

Using that the spectral norm is given by the square root of the scalar product and $\boldsymbol{U}$ having orthonormal columns, yields

$$\|\boldsymbol{w}\|^2 = \boldsymbol{w}^*\boldsymbol{w} = \boldsymbol{e}_L^*\boldsymbol{U}\boldsymbol{U}^*\boldsymbol{e}_L = \boldsymbol{e}_L^*\boldsymbol{U}\boldsymbol{U}^*\boldsymbol{U}\boldsymbol{U}^*\boldsymbol{e}_L = \|\boldsymbol{U}\boldsymbol{U}^*\boldsymbol{e}_L\|^2. \tag{4.3.11}$$

By construction, the columns of $\boldsymbol{U}$ form an orthonormal basis for range$(\boldsymbol{A}_L^\top)$, i.e. $\boldsymbol{U}\boldsymbol{U}^*$ is an orthogonal projection into range$(\boldsymbol{A}_L^\top)$. The orthogonal projection of a vector into a subspace provides a vector in this space with smallest angle, i.e. largest scalar product. Hence, we can continue (4.3.11) with

$$\|\boldsymbol{U}\boldsymbol{U}^*\boldsymbol{e}_L\|^2 = \max_{\substack{\boldsymbol{y} \in \text{range}(\boldsymbol{A}_L^\top) \\ \boldsymbol{y} \neq 0}} \frac{|\boldsymbol{e}_L^*\boldsymbol{y}|^2}{\|\boldsymbol{y}\|^2} = \max_{\substack{\boldsymbol{y} \in \text{range}(\boldsymbol{A}_L^\top) \\ \boldsymbol{y} \neq 0}} \frac{|(\boldsymbol{y})_L|^2}{\|\boldsymbol{y}\|^2}. \tag{4.3.12}$$

Now, we show that each such vector $\boldsymbol{y}$ is the Fourier coefficient vector of some specific discrete measure with support in the nodes set Ω. Therefore, we can further bound (4.3.12) in terms of the supremum over discrete measures. Let $\boldsymbol{a}_j$ be the j-th column of $\boldsymbol{A}_L^\top$. Then for each $\boldsymbol{y} \in \text{range}(\boldsymbol{A}_L^\top)$ there exists a unique (rank$(\boldsymbol{A}_L)$ is full) coefficient vector $\boldsymbol{c} = (c_1, \ldots, c_M)^\top \in \mathbb{C}^M$ such that $\boldsymbol{y} = \sum_{j=1}^M c_j\boldsymbol{a}_j$. We define the discrete complex measure

$$\mu := \sum_{j=1}^M c_j \mathrm{e}^{2\pi i L t_j} \delta_{t_j}, \tag{4.3.13}$$

with supp$(\mu) \subset \Omega$. Its Fourier coefficients, Definition 2.3.6, are given by

$$\widehat{\mu}(k) = \int_0^1 \mathrm{e}^{-2\pi i k t}\, d\mu(t) = \sum_{j=1}^M c_j \mathrm{e}^{2\pi i L t_j} \mathrm{e}^{-2\pi i k t_j} = \sum_{j=1}^M c_j \left(\boldsymbol{a}_j\right)_{L-k} = (\boldsymbol{y})_{L-k}.$$

Reading this chain of equations from right to left shows that for each vector $\boldsymbol{y} \in \operatorname{range}(\boldsymbol{A}_L^\top)$ there exists such a discrete measure of which the first L Fourier coefficients are equal to its entries (up to the inverse ordering). Thus, the so defined map from discrete measures to vectors in $\mathbb{C}^L$ is surjective and we can bound (4.3.12) in terms of measures and combine it with (4.3.11) to obtain

$$\|\boldsymbol{w}\|^2 \leq \sup_{\operatorname{supp}(\mu)\subset\Omega} \frac{|\widehat{\mu}(0)|^2}{\sum_{k=0}^{L-1} |\widehat{\mu}(k)|^2}.$$

□

Before we proceed with the main theorem, we state two lemmas which will be useful in the proof.

Lemma 4.3.9 ([74, Prop. 4])**.**
Let μ be an atomic measure supported on the node set Ω. If $f\colon \mathbb{T} \to \mathbb{C}$ is a continuous function with $f(t) = 1$ for all $t \in \Omega$ and $\operatorname{supp}(\widehat{f}) \subset \Xi \subset \mathbb{Z}$, then

$$|\widehat{\mu}(0)| \leq \|f\|_{L^2(\mathbb{T})} \left(\sum_{k\in\Xi} |\widehat{\mu}(k)|^2 \right)^{\frac{1}{2}}.$$

Proof. Let $\mu = \sum_{j=1}^M c_j \delta_{t_j}$ and $f(t) = \sum_{k\in\Xi} \widehat{f}_k \mathrm{e}^{2\pi i k t}$. Then applying the Cauchy–Schwarz inequality and Parseval's identity leads to

$$\begin{aligned} |\widehat{\mu}(0)| &= \left| \sum_{j=1}^M c_j \right| = \left| \sum_{j=1}^M c_j \overline{f(t_j)} \right| = \left| \int_{\mathbb{T}} \overline{f}\, d\mu \right| = \left| \sum_{k\in\Xi} \overline{\widehat{f}_k} \int_{\mathbb{T}} \mathrm{e}^{-2\pi i k t}\, d\mu \right| \\ &= \left| \sum_{k\in\Xi} \overline{\widehat{f}(k)} \widehat{\mu}(k) \right| \leq \left\| \widehat{\boldsymbol{f}} \right\| \left(\sum_{k\in\Xi} |\widehat{\mu}(k)|^2 \right)^{\frac{1}{2}} = \|f\|_{L^2(\mathbb{T})} \left(\sum_{k\in\Xi} |\widehat{\mu}(k)|^2 \right)^{\frac{1}{2}}. \end{aligned}$$

□

Lemma 4.3.10 (One-interpolating function, [74, pp. 10])**.**
Let $\Omega = \{t_1, \ldots, t_M\} \subset \mathbb{T}$ be a set of distinct nodes and $L \in \mathbb{N}, L > M$. Then there exists an $f \in C(\mathbb{T})$ with $\operatorname{supp}(\widehat{f}) = \{0, \ldots, L-1\}$ and satisfying the interpolation condition $f(t_j) = 1, j = 1, \ldots, M$, such that

$$\|f\|_{L^2(\mathbb{T})} \leq 1 - 4^{-M}.$$

Proof. Consider the trigonometric polynomial $p\colon \mathbb{T} \to \mathbb{C}$,

$$p(t) := (-1)^M \prod_{j=1}^M \left(\mathrm{e}^{2\pi i (t - t_j)} - 1 \right),$$

vanishing at the nodes in Ω. By multiplying out the expression, we see

$$p(t) = \sum_{k=0}^M \widehat{p}(k) \mathrm{e}^{2\pi i k t},$$

where $\widehat{p}(0) = 1$ and since $L > M$, we have $\operatorname{supp}(\widehat{p}) \subset \{0, \dots, M\} \subset \{0, \dots, L-1\}$. Furthermore, since $\left|\mathrm{e}^{2\pi i t} - \mathrm{e}^{2\pi i t_j}\right| \leq 2$, we have the uniform bound

$$\|p\|_{\mathcal{C}(\mathbb{T})} \leq \sup_{t \in \mathbb{T}} \prod_{j=1}^{M} \left|\mathrm{e}^{2\pi i t} - \mathrm{e}^{2\pi i t_j}\right| \leq 2^M. \tag{4.3.14}$$

Let $a > 0$ be a parameter. Now, the function f defined as

$$f(t) := 1 - ap(t)$$

has the interpolation property and its Fourier coefficients are given by $\widehat{f}(0) = 1 - a$, $\widehat{f}(k) = -a\widehat{p}(k), k = 1, \dots, M$ and $\widehat{f}(k) = 0$, else, i.e. its discrete Fourier transform has the right support. The Hölder inequality yields $\|p\|_{L^2(\mathbb{T})} \leq \|p\|_{\mathcal{C}(\mathbb{T})}$ and therefore, by Parseval's identity, we obtain

$$\begin{aligned}\|f\|_{L^2(\mathbb{T})}^2 &= \sum_{k=0}^{M} \left|\widehat{f}(k)\right|^2 = (1-a)^2 + a^2 \sum_{k=1}^{M} |\widehat{p}(k)|^2 = (1-a)^2 - a^2 |\widehat{p}(0)| + a^2 \sum_{k=0}^{M} |\widehat{p}(k)|^2 \\ &= (1-a)^2 - a^2 + a^2 \|p\|_{L^2(\mathbb{T})}^2 \leq 1 - 2a + a^2 \|p\|_{\mathcal{C}(\mathbb{T})}^2 .\end{aligned}$$

This upper bound is quadratic in a and minimized by $a = \|p\|_{\mathcal{C}(\mathbb{T})}^{-2}$. Using this special value for a and (4.3.14) yields

$$\|f\|_{L^2(\mathbb{T})}^2 \leq 1 - 4^{-M}.$$

□

Now, we can proof the main theorem of this section, which provides three different lower bounds for the smallest singular value of the truncated singular vector matrix $\boldsymbol{U}_{\uparrow}$. We see that one bound can be obtained directly as done in [5]. The remaining two bounds can be found in [74]. One is in terms of the smallest and largest singular value of corresponding Vandermonde matrices and the second depends only on the number of nodes M. We provide the full proof in order to be able to give some remarks afterwards. Comparisons of these bounds are give in the next section.

Theorem 4.3.11 (Lower bound on $\sigma_{\min}(\boldsymbol{U}_{\uparrow})$, [5, 74])**.**
Let matrices be given as in Section 4.2. Then we have as lower bounds for the smallest singular value of $\boldsymbol{U}_{\uparrow}$

$$\sigma_{\min}(\boldsymbol{U}_{\uparrow}) \geq \max\left\{ \frac{\sigma_{\min}(\boldsymbol{A}_{L-1})}{\sigma_{\max}(\boldsymbol{A}_L)}, \sqrt{1 - \frac{M}{\sigma_{\min}(\boldsymbol{A}_L)^2}}, 2^{-M} \right\}.$$

Proof. Analogous to [5, p. 202], the first bound can be established by using Lemma 2.1.16 and (4.3.7) in

$$\sigma_{\min}(\boldsymbol{U}_{\uparrow}) = \sigma_{\min}(\boldsymbol{A}_{\uparrow}\boldsymbol{P}) \geq \sigma_{\min}(\boldsymbol{A}_{\uparrow})\, \sigma_{\min}(\boldsymbol{P}) = \sigma_{\min}(\boldsymbol{A}_{L-1})\, \sigma_{\min}(\boldsymbol{P}) \geq \frac{\sigma_{\min}(\boldsymbol{A}_{L-1})}{\sigma_{\max}(\boldsymbol{A}_L)}.$$

The proof of the remaining bounds is from [74]. By Lemma 4.3.8, we have a lower bound in terms of

$$\sup_{\operatorname{supp}(\mu) \subset \Omega} \frac{|\widehat{\mu}(0)|^2}{\sum_{k=0}^{L-1} |\widehat{\mu}(k)|^2}. \tag{4.3.15}$$

Let $\mu = \sum_{j=1}^{M} c_j \delta_{t_j}$ be a measure with coefficient vector $\boldsymbol{c} = (c_1, \dots, c_M)^\top \in \mathbb{C}^M$ and support in Ω. All measures over which the supremum is taken can be written like that. Then its Fourier coefficients are given by

$$\widehat{\mu}(k) = \sum_{j=1}^{M} c_j \mathrm{e}^{-2\pi i k t_j} = (\boldsymbol{A}_L^* \boldsymbol{c})_k \,,$$

for $k = 0, \dots, L-1$. Therefore, by the Cauchy–Schwarz inequality it holds for the nominator in (4.3.15)

$$|\widehat{\mu}(0)|^2 = \left| \sum_{j=1}^{M} c_j \right|^2 \leq \|\boldsymbol{c}\|_1^2 \leq (\sqrt{M}\,\|\boldsymbol{c}\|)^2 = M\,\|\boldsymbol{c}\|^2 \,.$$

Using Theorem 2.1.15 and that singular values are invariant under transposing, we obtain for the denominator in (4.3.15)

$$\sum_{k=0}^{L-1} |\widehat{\mu}(k)|^2 = \left\| \left(\widehat{\mu}(k)\right)_{k=0}^{L-1} \right\|^2 = \|\boldsymbol{A}_L^* \boldsymbol{c}\|^2 = \boldsymbol{c}^* \boldsymbol{A}_L \boldsymbol{A}_L^* \boldsymbol{c} = \frac{\boldsymbol{c}^* \boldsymbol{A}_L \boldsymbol{A}_L^* \boldsymbol{c}}{\boldsymbol{c}^* \boldsymbol{c}} \|\boldsymbol{c}\|^2 \geq \sigma_{\min}(\boldsymbol{A}_L)\,\|\boldsymbol{c}\|^2 \,.$$

Combining the bounds for nominator and denominator, yields a bound independent on the respective measure, so that the supremum can be omitted. This yields together with Lemma 4.3.8

$$\sigma_{\min}(\boldsymbol{U}_\uparrow) \geq \sqrt{1 - \frac{M}{\sigma_{\min}(\boldsymbol{A}_L)^2}}$$

the second bound from the theorem.

The proof of the third bound is in the beginning analogous to that of the second, but instead of bounding (4.3.15) directly, we apply Lemma 4.3.9. Let $\Xi = \{0, \dots, L-1\}$ and we choose $f \in C(\mathbb{T})$ as in Lemma 4.3.10. Then, for every discrete measure μ with $\operatorname{supp}(\mu) \subset \Omega$, we have $|\widehat{\mu}(0)|^2 \leq \|f\|_{L^2(\mathbb{T})}^2 \sum_{k=0}^{L-1} |\widehat{\mu}(k)|^2$. We use this in (4.3.15) and get

$$\sup_{\operatorname{supp}(\mu) \subset \Omega} \frac{|\widehat{\mu}(0)|^2}{\sum_{k=0}^{L-1} |\widehat{\mu}(k)|^2} \leq \|f\|_{L^2(\mathbb{T})}^2 \,. \tag{4.3.16}$$

Now Lemma 4.3.10 provides the bound

$$\|f\|_{L^2(\mathbb{T})}^2 \leq 1 - 4^{-M}$$

uniformly in all node sets Ω of cardinality M. Together with (4.3.16) and Lemma 4.3.8, we obtain the third lower bound

$$\sigma_{\min}(\boldsymbol{U}_\uparrow) \geq \sqrt{1 - \|f\|_{L^2(\mathbb{T})}^2} \geq \sqrt{1 - (1 - 4^{-M})} = 2^{-M}.$$

□

Remark 4.3.12.
In the proof of the third inequality, we used the interpolating function from Lemma 4.3.10, that has a minimal number of nonzero Fourier coefficients. One can ask whether it is possible to obtain better results by exploiting the whole support in Fourier domain. In order to give

a partial answer, let $f \in C(\mathbb{T})$*, such that* $f(t_j) = 1$*, for* $j = 1, \dots, M$*, and* $\operatorname{supp}(\widehat{f}) \subset \{0, \dots, L-1\}$*. This is equivalent to the system of equations* $f(t_j) = \sum_{k=0}^{L-1} \widehat{f}(k) \mathrm{e}^{2\pi i k t_j} = 1$ *for* $j = 1, \dots, M$*. Using the Vandermonde matrix* $\boldsymbol{A}_L$ *and denoting by* $\mathbb{1}_{M\times 1}$ *the vector of ones in* $\mathbb{C}^M$*, we can rewrite the equations to*

$$\boldsymbol{A}_L \widehat{\boldsymbol{f}} = \mathbb{1}_{M\times 1}. \tag{4.3.17}$$

Due to Parseval's identity, we have $\|f\|_{L^2(\mathbb{T})} = \left\|\widehat{\boldsymbol{f}}\right\|$*. Therefore, searching for a minimal* L^2*-norm function* f*, that satisfies the above properties, is equivalent to finding the best* 2*-norm solution to* (4.3.17)*. This can be done by multiplying* (4.3.17) *from the left with the Moore-Penrose pseudo inverse* $\boldsymbol{A}_L^\dagger$ *(see [58, Fact 4.23 and Fact 5.8]) to obtain*

$$\widehat{\boldsymbol{f}} = \boldsymbol{A}_L^\dagger \mathbb{1}_{M\times 1}.$$

Hence, a bound on the L^2*-norm of an optimal* f *is given by*

$$\|f\|_{L^2(\mathbb{T})} = \left\|\widehat{\boldsymbol{f}}\right\| = \left\|\boldsymbol{A}_L^\dagger \mathbb{1}_{M\times 1}\right\| \le \left\|\boldsymbol{A}_L^\dagger\right\| \|\mathbb{1}_{M\times 1}\| \le \frac{\sqrt{M}}{\sigma_{\min}(\boldsymbol{A}_L)}.$$

Together with (4.3.16) *and Lemma 4.3.8 this yields again the second bound of Theorem 4.3.11. To conclude, we see that improving bounds is only possible if the* L^2*-norm of the function* f *is directly controlled. Using geometric information about the node set* Ω*, e.g. being well separated or being a clustered node configuration, could be helpful, too. This is also mentioned in [74, p. 5].*

4.3.3 Comparison and error bounds

In Chapter 3, we studied bounds on the extremal singular values of Vandermonde matrices with clustered node configurations on the unit circle. In the following we use these bounds to refine the estimates for the ESPRIT algorithm. First of all, we apply them to the bounds for $\sigma_{\min}(\boldsymbol{U}_\uparrow)$ from Theorem 4.3.11 and decide which of them should be used afterwards in which situation.

Lemma 4.3.13 (Comparison of lower bounds for $\sigma_{\min}(\boldsymbol{U}_\uparrow)$)**.**
Using the different bounds from Chapter 3 in the well-separated and clustered nodes regimes for the singular values of the Vandermonde matrix $\boldsymbol{A}_L$ *or* $\boldsymbol{A}_{L-1}$*, we can further bound the ones from Theorem 4.3.11 and obtain the following table.*

bounds from Theorem 4.3.11	well-separated, $qL > 1$	clustered node configuration
$\sigma_{\min}(\boldsymbol{U}_\uparrow) \ge \frac{\sigma_{\min}(\boldsymbol{A}_{L-1})}{\sigma_{\max}(\boldsymbol{A}_L)} \ge$	$\sqrt{\frac{qL-1}{qL+1}}$	$\frac{1}{1.8\sqrt{2\lambda}} \sqrt{\frac{L-1}{L}} \left(\frac{q(L-1)}{2\mathrm{e}}\right)^{\lambda-1}$
$\sigma_{\min}(\boldsymbol{U}_\uparrow) \ge \sqrt{1 - \frac{M}{\sigma_{\min}(\boldsymbol{A}_L)^2}} \ge$	$\sqrt{1 - \frac{qL}{qL-1} \cdot \frac{M}{L}}$	$\sqrt{1 - \frac{3.24M}{L} \left(\frac{2\mathrm{e}}{qL}\right)^{2\lambda-2}}$
$\sigma_{\min}(\boldsymbol{U}_\uparrow) \ge 2^{-M} \ge$	2^{-M}	2^{-M}

Assumptions for the case of clustered node configurations are the same as the ones from Corollary 3.4.13. In case of well-separated nodes, clearly the first or second bound should be preferred, where the second bound is better only if L *is significantly larger than the number of nodes* M*. For clustered node configurations the third bound is most suitable when almost all*

nodes build a cluster, i.e. $\lambda \approx M$. Otherwise the first bound should be preferred. The second bound is non-trivial only when L is greater than M times a constant, that is exponential in λ and thus, not favorable for the situation with clustered node configurations. The exact conditions are stated in the proof.

Proof. We start with case of well separated nodes, i.e. we assume $q > 1/L$ and choose the known bounds in a way that re-normalization factors are minimized. By Theorem 3.1.7, we have $\sigma_{\max}(\boldsymbol{A}_L) \leq \sqrt{L + \frac{1}{q}}$ and $\sigma_{\min}(\boldsymbol{A}_{L-1}) \geq \sqrt{L - \frac{1}{q}}$. Combining both yields

$$\frac{\sigma_{\min}(\boldsymbol{A}_{L-1})}{\sigma_{\max}(\boldsymbol{A}_L)} \geq \frac{\sqrt{L - \frac{1}{q}}}{\sqrt{L + \frac{1}{q}}} = \sqrt{\frac{1 - \frac{1}{qL}}{1 + \frac{1}{qL}}} = \sqrt{\frac{qL - 1}{qL + 1}}.$$

Using the slightly weaker bound $\sigma_{\min}(\boldsymbol{A}_L) \geq \sqrt{L - \frac{1}{q}}$ yields

$$\sqrt{1 - \frac{M}{\sigma_{\min}(\boldsymbol{A}_L)^2}} \geq \sqrt{1 - \frac{M}{L(1 - \frac{1}{qL})}} = \sqrt{1 - \frac{qL}{qL - 1} \cdot \frac{M}{L}}.$$

This bound is non-trivial as long as $L > \frac{qL}{qL-1}M$. It is better (larger) than the first bound if and only if $1 - \frac{qL}{qL-1} \cdot \frac{M}{L} > \frac{qL-1}{qL+1}$, which is the case if and only if $\frac{2(qL-1)}{qL(qL+1)} > \frac{M}{L}$. This is equivalent to the condition

$$L > \frac{qL(qL+1)}{2(qL-1)}M.$$

The minimum of the factor before M is achieved for $qL \in [2, 3]$ and larger than 2.5. Now we consider clustered node configurations. Let the cluster separation $\Delta L \geq 4.4\lambda \left(\frac{2\mathrm{e}}{qL}\right)^{\frac{\lambda-1}{2}}$. From Corollary 3.4.13 we know that

$$\sigma_{\min}(\boldsymbol{A}_{L-1}) \geq \sqrt{L-1}\frac{(q(L-1))^{\lambda-1}}{1.8(2\mathrm{e})^{\lambda-1}}$$

and from Lemma 3.4.6 that $\sigma_{\max}(\boldsymbol{A}_L) \leq \sqrt{L}\sqrt{2\lambda}$. Therefore, we get

$$\frac{\sigma_{\min}(\boldsymbol{A}_{L-1})}{\sigma_{\max}(\boldsymbol{A}_L)} \geq \frac{1}{1.8\sqrt{2\lambda}}\sqrt{\frac{L-1}{L}}\left(\frac{q(L-1)}{2\mathrm{e}}\right)^{\lambda-1}.$$

The second bound is obtained by simply inserting $\sigma_{\min}(\boldsymbol{A}_L)^2 \geq L\frac{(qL)^{2\lambda-2}}{3.24(2\mathrm{e})^{2\lambda-2}}$ in

$$\sqrt{1 - \frac{M}{\sigma_{\min}(\boldsymbol{A}_L)^2}} \geq \sqrt{1 - \frac{3.24M}{L}\left(\frac{2\mathrm{e}}{qL}\right)^{2\lambda-2}}.$$

It is non-trivial as long as $L > 3.24M(\frac{2\mathrm{e}}{qL})^{2\lambda-2}$ and better than the first bound if

$$1 - \frac{3.24M}{L}\left(\frac{2\mathrm{e}}{qL}\right)^{2\lambda-2} \geq \frac{1}{1.8^2 \cdot 2\lambda}\left(\frac{L-1}{L}\right)^{2\lambda-1}\left(\frac{qL}{2\mathrm{e}}\right)^{2\lambda-2}.$$

This is equivalent to

$$L > 3.24M \left(\frac{2\mathrm{e}}{qL}\right)^{2\lambda-2} \left(1 - \frac{1}{1.8^2 \cdot 2\lambda} \left(\frac{L-1}{L}\right)^{2\lambda-1} \left(\frac{qL}{2\mathrm{e}}\right)^{2\lambda-2}\right)^{-1},$$

which is fulfilled if

$$L > 3.24M \left(\frac{2\mathrm{e}}{qL}\right)^{2\lambda-2} \left(1 - \frac{1}{6.48\lambda} \left(\frac{qL}{2\mathrm{e}}\right)^{2\lambda-2}\right)^{-1}$$

since $\frac{L-1}{L} \leq 1$ and $1.8^2 \cdot 2 = 6.48$.

Finally, we compare the first bound with 2^{-M}. In the well separated regime we directly see that the first bound should be preferred, since it is completely independent of M. In the situation of clustered nodes, the first bound has a factor $(2\mathrm{e})^{\lambda-1}$. From this, we already see that it can only be better than 2^{-M} if λ is much smaller than M. More precisely, we have

$$\frac{1}{1.8\sqrt{2\lambda}} \left(\frac{L-1}{L}\right)^{\lambda-\frac{1}{2}} \left(\frac{qL}{2\mathrm{e}}\right)^{\lambda-1} \leq 2^{-M}$$

as long as

$$\frac{1}{1.8\sqrt{2}} \left(\frac{qL}{2\mathrm{e}}\right)^{\lambda-1} \leq 2^{-M}.$$

After taking the logarithm on both sides, we have that 2^{-M} should be preferred to the first bound if

$$\lambda \geq M \frac{\log(2)}{\log(2\mathrm{e}(qL)^{-1})} - \frac{\log(1.8\sqrt{2})}{\log(2\mathrm{e}(qL)^{-1})} + 1.$$

□

Now we use the respective most suitable bounds from Lemma 4.3.13 (well-separated: first row; clustered nodes: first row for small clusters, third row for large clusters) in combination with Theorem 4.3.5 and our knowledge about the conditioning of Vandermonde matrices from Chapter 3 to state final results for the stability of the node reconstruction via the ESPRIT algorithm. In order to keep the results as comprehensible as possible, we choose as parameter for the ESPRIT algorithm $L = \lceil \frac{N}{2} \rceil$. This makes things easier, because if $N = 2k$ for some $k \in \mathbb{N}$, then

$$N - L + 1 = N - \left\lceil \frac{N}{2} \right\rceil + 1 = k + 1 \geq k = \left\lceil \frac{N}{2} \right\rceil = L$$

and if $N = 2k + 1$, then

$$N - L + 1 = N - \left\lceil \frac{N}{2} \right\rceil + 1 = 2k + 1 - k - 1 + 1 = k + 1 = \left\lceil \frac{N}{2} \right\rceil = L.$$

This means for all N, we have $N - L + 1 \geq L$ and therefore, by monotony of singular values (Lemma 2.1.18),

$$\sigma_{\min}(\boldsymbol{A}_{N-L+1}) \geq \sigma_{\min}(\boldsymbol{A}_L). \tag{4.3.18}$$

Hence, we only need lower bounds for $\sigma_{\min}(\boldsymbol{A}_L)$ from Chapter 3 to proof the following theorems.

Theorem 4.3.14 (Node reconstruction error for well-separated nodes)**.**
Let $L = \lceil \frac{N}{2} \rceil$ and let the node set $\Omega = \{t_1, \dots, t_M\} \subset \mathbb{T}$ consist of well-separated nodes, i.e. the minimal separation distance $q > 1/L$. Then for the node reconstruction in 4.1.1 with the ESPRIT algorithm, we have the following. If the norm of the error in the Hankel matrix satisfies

$$\|\boldsymbol{E}\| \leq \frac{L \cdot \alpha_{\min}}{4\sqrt{2}} \left(\frac{qL-1}{qL+1}\right)^{\frac{3}{2}},$$

then the matching distance between original nodes and reconstructed nodes fulfills

$$\mathrm{md}_{\mathbb{T}}(\Omega, \widetilde{\Omega}) \leq \frac{6(2M-1)}{L \cdot \alpha_{\min}} \left(\frac{qL+1}{qL-1}\right)^{\frac{5}{2}} \|\boldsymbol{E}\|.$$

Proof. We apply Theorem 4.3.5 and the first bound from Theorem 4.3.11 with its refinement from Lemma 4.3.13. Theorem 3.1.7 yields

$$\sigma_{\min}(\boldsymbol{A}_L) \geq \sqrt{L}\sqrt{\frac{qL-1}{qL}} \tag{4.3.19}$$

and Lemma 4.3.13

$$\sigma_{\min}(\boldsymbol{U}_{\uparrow}) \geq \sqrt{\frac{qL-1}{qL+1}}. \tag{4.3.20}$$

Theorem 4.3.5 needs

$$\|\boldsymbol{E}\| \leq \frac{\alpha_{\min}\, \sigma_{\min}(\boldsymbol{A}_L)\, \sigma_{\min}(\boldsymbol{A}_{N-L+1})\, \sigma_{\min}(\boldsymbol{U}_{\uparrow})}{4\sqrt{2}}, \tag{4.3.21}$$

which is by (4.3.19), (4.3.20) and use of (4.3.18) satisfied, if

$$\|\boldsymbol{E}\| \leq \frac{\alpha_{\min} \cdot L(qL-1)\sqrt{qL-1}}{4\sqrt{2}qL\sqrt{\tau+1}}.$$

This in turn holds if the assumption on $\|\boldsymbol{E}\|$ from this theorem are fulfilled. Here we loose a little in order to keep the expression simple. Now we apply Theorem 4.3.5 together with (4.3.18) and get

$$\mathrm{md}_{\mathbb{T}}(\Omega, \widetilde{\Omega}) \leq \frac{6(2M-1)\,\|\boldsymbol{A}_L\|}{\alpha_{\min}\, \sigma_{\min}(\boldsymbol{A}_L)^3\, \sigma_{\min}(\boldsymbol{U}_{\uparrow})^2} \|\boldsymbol{E}\|. \tag{4.3.22}$$

From Theorem 3.1.7, we have $\|\boldsymbol{A}_L\| \leq \sqrt{L}\sqrt{\frac{qL+1}{qL}}$ and hence,

$$\begin{aligned}
\mathrm{md}_{\mathbb{T}}(\Omega, \widetilde{\Omega}) &\leq \frac{6(2M-1)\sqrt{L}\sqrt{qL+1}(qL)^{\frac{3}{2}}(qL+1)}{\alpha_{\min} \cdot L^{\frac{3}{2}}\sqrt{qL}(qL-1)^{\frac{3}{2}}(qL-1)} \|\boldsymbol{E}\| \\
&\leq \frac{6(2M-1)\sqrt{L}\sqrt{qL+1}(qL+1)^2}{\alpha_{\min} \cdot L^{\frac{3}{2}}(qL-1)^{\frac{3}{2}}(qL-1)} \|\boldsymbol{E}\|.
\end{aligned}$$

This is after rearranging the result. Notice that we weaken the estimate in the last inequality slightly in order to simplify the expression. □

In case of clustered node configurations, we have the following theorems for the node reconstruction with the ESPRIT algorithm.

Theorem 4.3.15 (Node reconstruction error for node sets with small cluster sizes)**.**
Let $L = \lceil \frac{N}{2} \rceil$ and the node set $\Omega = \{t_1, \dots, t_M\} \subset \mathbb{T}$ be a clustered node configuration with minimal separation distance q, maximal cluster size λ and cluster separation

$$\Delta L \geq 4.4\lambda \left(\frac{2\mathrm{e}}{qL}\right)^{\frac{\lambda-1}{2}}.$$

Assume $L \geq 8\lambda$. Then for the node reconstruction in (4.1.1) with the ESPRIT algorithm we have the following. If the perturbation of the Hankel matrix satisfies

$$\|\boldsymbol{E}\| \leq \frac{L \cdot \alpha_{\min}}{47\sqrt{\lambda}} \left(\frac{qL}{2\mathrm{e}}\right)^{3\lambda-3} \left(\frac{L-1}{L}\right)^{\lambda-\frac{1}{2}},$$

then the matching distance between original nodes and reconstructed nodes fulfills

$$\mathrm{md}_{\mathbb{T}}(\Omega, \widetilde{\Omega}) \leq \frac{294\lambda^{\frac{3}{2}}(2M-1)}{L \cdot \alpha_{\min}} \left(\frac{2\mathrm{e}}{qL}\right)^{5\lambda-5} \left(\frac{L}{L-1}\right)^{2\lambda-1} \|\boldsymbol{E}\|.$$

Proof. The proof is analogous to that of Theorem 4.3.14 with the difference that we use the lower bounds on $\sigma_{\min}(\boldsymbol{A}_L)$ and upper bounds on $\sigma_{\max}(\boldsymbol{A}_L)$ in the situation of clustered node configurations from Corollary 3.4.13 and Lemma 3.4.6. The assumption on the cluster separation Δ in Corollary 3.4.13 are the same as we demand in this theorem. Thus it provides the bounds

$$\sigma_{\min}(\boldsymbol{A}_L) \geq \sqrt{L} \frac{1}{1.8} \left(\frac{qL}{2\mathrm{e}}\right)^{\lambda-1}$$

For clustered node configurations with relative small cluster sizes, we use the first bound from Theorem 4.3.11 as analyzed in Lemma 4.3.13, i.e.

$$\sigma_{\min}(\boldsymbol{U}_{\uparrow}) \geq \frac{1}{1.8\sqrt{2\lambda}} \left(\frac{L-1}{L}\right)^{\lambda-\frac{1}{2}} \left(\frac{qL}{2\mathrm{e}}\right)^{\lambda-1}.$$

Then the assumptions on the error matrix $\boldsymbol{E}$ in Theorem 4.3.14 (recapped in (4.3.21)) are fulfilled, by additionally using (4.3.18), if

$$\|\boldsymbol{E}\| \leq \frac{\alpha_{\min}}{4\sqrt{2}} \cdot \frac{L}{1.8^2} \left(\frac{qL}{2\mathrm{e}}\right)^{2\lambda-2} \cdot \frac{1}{1.8\sqrt{2\lambda}} \left(\frac{L-1}{L}\right)^{\lambda-\frac{1}{2}} \left(\frac{qL}{2\mathrm{e}}\right)^{\lambda-1}.$$

After a slight estimate of the constant, this is satisfied by the assumption of this theorem. Now, applying Theorem 4.3.5 (recapped in (4.3.22) with (4.3.18) already utilized), using the upper bound

$$\sigma_{\max}(\boldsymbol{A}_L) \leq \sqrt{L} \cdot \sqrt{2\lambda}$$

from Theorem 3.1.7 and the above lower bounds for the smallest singular values yield

$$\mathrm{md}_{\mathbb{T}}(\Omega, \widetilde{\Omega}) \leq \frac{6(2M-1)\sqrt{L} \cdot \sqrt{2\lambda}}{\alpha_{\min}} \cdot 1.8^2 \cdot 2\lambda \left(\frac{L}{L-1}\right)^{2\lambda-1} \left(\frac{2\mathrm{e}}{qL}\right)^{2\lambda-2} \cdot \frac{1.8^3}{L^{\frac{3}{2}}} \left(\frac{2\mathrm{e}}{qL}\right)^{3\lambda-3} \|\boldsymbol{E}\|.$$

Estimating the constant and rearranging finishes the proof. □

Theorem 4.3.16 (Node reconstruction error for node sets with large cluster sizes)**.**
Let $L = \lceil \frac{N}{2} \rceil$ and the node set $\Omega = \{t_1, \dots, t_M\} \subset \mathbb{T}$ be a clustered node configuration with minimal separation distance q, maximal cluster size λ and cluster separation

$$\Delta L \geq 4.4\lambda \left(\frac{2\mathrm{e}}{qL}\right)^{\frac{\lambda-1}{2}}.$$

Then for the node reconstruction in (4.1.1) *with the ESPRIT algorithm we have the following. If the norm of the error in the Hankel matrix satisfies*

$$\|\boldsymbol{E}\| \leq \frac{L \cdot \alpha_{\min}}{19 \cdot 2^M} \left(\frac{qL}{2\mathrm{e}}\right)^{2\lambda-2},$$

then the matching distance between original nodes and reconstructed nodes fulfills

$$\mathrm{md}_{\mathbb{T}}(\Omega, \widetilde{\Omega}) \leq \frac{46 \cdot 4^M \sqrt{\lambda}(2M-1)}{L \cdot \alpha_{\min}} \left(\frac{2\mathrm{e}}{qL}\right)^{3\lambda-3} \|\boldsymbol{E}\|.$$

Proof. The proof is analogous to that of Theorem 4.3.15. The only difference is that we use the third bound

$$\sigma_{\min}(\boldsymbol{U}_{\uparrow}) \geq 2^{-M}.$$

from Theorem 4.3.11 is used. The condition on $\boldsymbol{E}$ in Theorem 4.3.5 is then satisfied if

$$\|\boldsymbol{E}\| \leq \frac{\alpha_{\min}}{4\sqrt{2}} \cdot \frac{L}{1.8^2} \left(\frac{qL}{2\mathrm{e}}\right)^{2\lambda-2} \cdot 2^{-M},$$

which is in turn fulfilled by the assumption of this theorem. Furthermore, applying Theorem 4.3.5 and using above bounds give

$$\mathrm{md}_{\mathbb{T}}(\Omega, \widetilde{\Omega}) \leq \frac{6(2M-1)\sqrt{L} \cdot \sqrt{2\lambda}}{\alpha_{\min}} \cdot 4^M \cdot \frac{1.8^3}{L^{\frac{3}{2}}} \left(\frac{2\mathrm{e}}{qL}\right)^{3\lambda-3} \|\boldsymbol{E}\|.$$

Finally, bounding the constant and rearranging completes the proof. □

Remark 4.3.17 (Independence of overall number of nodes)**.**
In contrast to [74, Lem. 5, p. 13], we used advanced properties of the principal vector matrices by Lemma 2.2.8. Therefore, we were able to directly bound the spectral norm of the difference matrix (not making a detour over the Frobenius norm) and obtained a bound independent of the number of nodes M. Hence, in Theorems 4.3.14 and 4.3.15 the dependence on M only appears in the factor $(2M-1)$, which is due to application of Theorem 2.2.3. In fact, looking into [104, Ch. 4, Thm. 3.3], from which we cited the theorem, we see that the factor comes from bounding the matching distance. More precisely, the bound on the matching distance is established by taking a bound on the so called spectral variation *times an additional factor $(2M-1)$. The spectral variation of the matrix $\widetilde{\boldsymbol{M}} \in \mathbb{C}^{m \times m}$ (with eigenvalues $\widetilde{\lambda}_1, \dots, \widetilde{\lambda}_m$) with respect to the matrix $\boldsymbol{M} \in \mathbb{C}^{m \times m}$ (with eigenvalues $\lambda_1, \dots, \lambda_m$) is defined by*

$$\mathrm{sv}_{\boldsymbol{M}}(\widetilde{\boldsymbol{M}}) := \max_{1 \leq j \leq m} \min_{1 \leq k \leq m} \left|\widetilde{\lambda}_j - \lambda_k\right|.$$

This means the bounds in Theorems 4.3.14 and 4.3.15 for the matching distance without the factor provide M-independent bounds on the spectral variation. However, in context of Theorems 4.3.14 and 4.3.15, a bound on the spectral variation only allows statements like, for each reconstructed node $\widetilde{z} \in \widetilde{\Lambda}$ there exists a $z_j \in \Lambda$ such that the bound on the spectral variation holds as upper bound for $|\widetilde{z} - z_j|$.

4.4 Stability of the coefficient reconstruction

The last step is to analyze the conditioning of the coefficient reconstruction by means of the least squares problem (4.2.7). The difficulty is here that we have to deal not only with a perturbed right hand side of the least squares system, i.e. the perturbed data vector $\widetilde{\boldsymbol{\mathcal{E}}}$, but also with a perturbed system matrix, more precisely the Vandermonde matrix corresponding to the reconstructed nodes. We follow the suggestion of [94, p. 642] and use a perturbation theorem by Wedin which can be found in [53, Thm. 20.1]. The results are for real matrices and vectors, but can be applied to the complex case easily. The theorem stated there provides a bound on the error between the solution $\boldsymbol{x}_0$ to a least squares problem $\min_x \|\boldsymbol{M}\boldsymbol{x}-\boldsymbol{b}\|$ (for some full rank matrix $\boldsymbol{M}\in\mathbb{C}^{m\times n}, n\leq m$ and a vector $\boldsymbol{b}\in\mathbb{C}^m$) and the solution $\boldsymbol{y}_0$ to the perturbed problem $\min_y \left\|\widetilde{\boldsymbol{M}}\boldsymbol{y}-\widetilde{\boldsymbol{b}}\right\|$. There, the vector $\boldsymbol{b}$ does not have to lie in the column space of $\boldsymbol{M}$ so that the residual $\boldsymbol{M}\boldsymbol{x}-\boldsymbol{b}$ could be nonzero. The situation in (4.2.7) is different. By assumption we have no residual in the exact case. Therefore, we adapt the proof of [53, Thm. 20.1] for our situation in the next lemma.

Lemma 4.4.1 (Least squares Perturbation bound, cf. [53, Thm. 20.1])**.**
Let $n,m\in\mathbb{N}_+$, $n\leq m$ and $\boldsymbol{M},\widetilde{\boldsymbol{M}}\in\mathbb{C}^{m\times n}$. Assume $\operatorname{rank}(\boldsymbol{M})=\operatorname{rank}(\widetilde{\boldsymbol{M}})=n$ *and hence both matrices have full rank. Let $\boldsymbol{b}=\boldsymbol{M}\boldsymbol{a}$ for some $\boldsymbol{a}\in\mathbb{C}^n$ and let $\widetilde{\boldsymbol{a}}\in\mathbb{C}^n$ be the solution to the least squares problem*

$$\min_{\boldsymbol{x}\in\mathbb{C}^n}\left\|\widetilde{\boldsymbol{M}}\boldsymbol{x}-\widetilde{\boldsymbol{b}}\right\|,$$

where $\widetilde{\boldsymbol{b}}\in\mathbb{C}^m$. If $\left\|\boldsymbol{M}-\widetilde{\boldsymbol{M}}\right\|<\frac{\sigma_{\min}(\boldsymbol{M})}{2}$, then the relative error between $\boldsymbol{a}$ and $\widetilde{\boldsymbol{a}}$ is given by

$$\frac{\|\boldsymbol{a}-\widetilde{\boldsymbol{a}}\|}{\|\boldsymbol{a}\|}\leq 2\operatorname{cond}(\boldsymbol{M})\left(\frac{\left\|\boldsymbol{b}-\widetilde{\boldsymbol{b}}\right\|}{\|\boldsymbol{b}\|}+\frac{\left\|\boldsymbol{M}-\widetilde{\boldsymbol{M}}\right\|}{\|\boldsymbol{M}\|}\right).$$

Proof. First of all, since $\boldsymbol{M}$ and $\widetilde{\boldsymbol{M}}$ have full rank, we have $\boldsymbol{M}^\dagger\boldsymbol{M}=\boldsymbol{I}_n=\widetilde{\boldsymbol{M}}^\dagger\widetilde{\boldsymbol{M}}$ and hence $\boldsymbol{a}=\boldsymbol{M}^\dagger\boldsymbol{b}$ and $\widetilde{\boldsymbol{a}}=\widetilde{\boldsymbol{M}}^\dagger\widetilde{\boldsymbol{b}}$. In particular, by assumption the residue $\boldsymbol{b}-\boldsymbol{M}\boldsymbol{a}=0$ and therefore, we follow the steps in [53, p. 401] with the according adjustments. Direct calculation shows

$$\begin{aligned}
\widetilde{\boldsymbol{a}}-\boldsymbol{a}&=\widetilde{\boldsymbol{M}}^\dagger\widetilde{\boldsymbol{b}}-\boldsymbol{a}=\widetilde{\boldsymbol{M}}^\dagger(\boldsymbol{M}\boldsymbol{a}+(\widetilde{\boldsymbol{b}}-\boldsymbol{b}))-\boldsymbol{a}\\
&=\widetilde{\boldsymbol{M}}^\dagger(\widetilde{\boldsymbol{M}}\boldsymbol{a}-(\widetilde{\boldsymbol{M}}-\boldsymbol{M})\boldsymbol{a}+(\widetilde{\boldsymbol{b}}-\boldsymbol{b}))-\boldsymbol{a}\\
&=\widetilde{\boldsymbol{M}}^\dagger((\widetilde{\boldsymbol{b}}-\boldsymbol{b})-(\widetilde{\boldsymbol{M}}-\boldsymbol{M})\boldsymbol{a})-(\boldsymbol{I}_n-\widetilde{\boldsymbol{M}}^\dagger\widetilde{\boldsymbol{M}})\boldsymbol{a}\\
&=\widetilde{\boldsymbol{M}}^\dagger((\widetilde{\boldsymbol{b}}-\boldsymbol{b})-(\widetilde{\boldsymbol{M}}-\boldsymbol{M})\boldsymbol{a}).
\end{aligned}$$

The full rank combined with the assumption $\left\|\boldsymbol{M}-\widetilde{\boldsymbol{M}}\right\|\leq\frac{\sigma_{\min}(\boldsymbol{M})}{2}$ and Lemma 2.1.16 yields

$$\left\|\widetilde{\boldsymbol{M}}^\dagger\right\|=\frac{1}{\sigma_{\min}(\widetilde{\boldsymbol{M}})}\leq\frac{1}{\sigma_{\min}(\boldsymbol{M})-\left\|\boldsymbol{M}-\widetilde{\boldsymbol{M}}\right\|}\leq\frac{2}{\sigma_{\min}(\boldsymbol{M})}.$$

Using this in the above identity and taking norms leads to

$$\|\boldsymbol{a}-\widetilde{\boldsymbol{a}}\|\leq\left\|\widetilde{\boldsymbol{M}}^\dagger\right\|\left(\left\|\boldsymbol{b}-\widetilde{\boldsymbol{b}}\right\|+\left\|\boldsymbol{M}-\widetilde{\boldsymbol{M}}\right\|\|\boldsymbol{a}\|\right).$$

We divide by $\|\boldsymbol{a}\|$ and since $\|\boldsymbol{b}\| \leq \|\boldsymbol{M}\| \|\boldsymbol{a}\|$, we have $\frac{1}{\|\boldsymbol{a}\|} \leq \frac{\|\boldsymbol{M}\|}{\|\boldsymbol{b}\|}$ and thus proceed with

$$\begin{aligned}\frac{\|\boldsymbol{a}-\widetilde{\boldsymbol{a}}\|}{\|\boldsymbol{a}\|} &\leq \left\|\widetilde{\boldsymbol{M}}^{\dagger}\right\| \left(\frac{\left\|\boldsymbol{b}-\widetilde{\boldsymbol{b}}\right\|}{\|\boldsymbol{a}\|} + \left\|\boldsymbol{M}-\widetilde{\boldsymbol{M}}\right\|\right) \leq \left\|\widetilde{\boldsymbol{M}}^{\dagger}\right\| \|\boldsymbol{M}\| \left(\frac{\left\|\boldsymbol{b}-\widetilde{\boldsymbol{b}}\right\|}{\|\boldsymbol{b}\|} + \frac{\left\|\boldsymbol{M}-\widetilde{\boldsymbol{M}}\right\|}{\|\boldsymbol{M}\|}\right) \\ &\leq \frac{2\|\boldsymbol{M}\|}{\sigma_{\min}(\boldsymbol{M})} \left(\frac{\left\|\boldsymbol{b}-\widetilde{\boldsymbol{b}}\right\|}{\|\boldsymbol{b}\|} + \frac{\left\|\boldsymbol{M}-\widetilde{\boldsymbol{M}}\right\|}{\|\boldsymbol{M}\|}\right) = 2\operatorname{cond}(\boldsymbol{M}) \left(\frac{\left\|\boldsymbol{b}-\widetilde{\boldsymbol{b}}\right\|}{\|\boldsymbol{b}\|} + \frac{\left\|\boldsymbol{M}-\widetilde{\boldsymbol{M}}\right\|}{\|\boldsymbol{M}\|}\right).\end{aligned}$$

□

Theorem 4.4.2 (Coefficient reconstruction error for well-separated nodes)**.**
Define $\alpha_{\max} := \max_{j=1,\dots,M} |\alpha_j|$. If Ω is a set of well-separated nodes with minimum separation distance $q > \frac{2}{N}$ and the error in the samples satisfies

$$\left\|\boldsymbol{\mathcal{E}}-\widetilde{\boldsymbol{\mathcal{E}}}\right\| \leq \frac{\alpha_{\min}}{34\sqrt{2}\pi\sqrt{MN}(2M-1)} \left(\frac{qN-2}{qN+2}\right)^{\frac{5}{2}},$$

then the relative error for the reconstruction of the exponential sum (4.1.1) *with the ESPRIT algorithm (paramter $L = \lceil \frac{N}{2} \rceil$) satisfies*

$$\frac{\|\boldsymbol{\alpha}-\widetilde{\boldsymbol{\alpha}}\|}{\|\boldsymbol{\alpha}\|} \leq \sqrt{3}\left(1 + 17\pi MN(2M-1)\frac{\alpha_{\max}}{\alpha_{\min}}\left(\frac{qN+2}{qN-2}\right)^{\frac{5}{2}}\right)\frac{\left\|\boldsymbol{\mathcal{E}}-\widetilde{\boldsymbol{\mathcal{E}}}\right\|}{\|\boldsymbol{\mathcal{E}}\|}.$$

Proof. In order to reconstruct the coefficients of the exponential sum $\mathcal{E}$, we apply Lemma 4.4.1 to the equation $\boldsymbol{A}_N^{\top}\boldsymbol{\alpha} = \boldsymbol{\mathcal{E}}$ and the least squares problem $\min_{\boldsymbol{x}\in\mathbb{C}^M} \left\|\widetilde{\boldsymbol{A}}_N^{\top}\widetilde{\boldsymbol{\alpha}} - \widetilde{\boldsymbol{\mathcal{E}}}\right\|$, where $\boldsymbol{A}_N$ and $\widetilde{\boldsymbol{A}}_N$ are the Vandermonde matrices corresponding to the node sets Ω and $\widetilde{\Omega}$, respectively. For this, we firstly bound the norm of the difference between the Vandermonde matrices. Two nodes $z = \mathrm{e}^{2\pi i t}, \widetilde{z} = \mathrm{e}^{2\pi i \widetilde{t}} \in \mathbb{C}$ on the unit circle satisfy the inequality $|z-\widetilde{z}| = 2\sin(\pi\left|t-\widetilde{t}\right|_{\mathbb{T}}) \leq 2\pi\left|t-\widetilde{t}\right|_{\mathbb{T}}$. Hence, it holds, for $k \in \mathbb{N}$,

$$\left|z_j^k - \widetilde{z}_j^k\right| \leq 2\pi k\left|t_j - \widetilde{t}_j\right|_{\mathbb{T}}, \quad j = 1,\dots,M.$$

This is only useful as long as $\left|t-\widetilde{t}\right|_{\mathbb{T}}$ is small enough since the trivial bound $\left|z_j^k - \widetilde{z}_j^k\right| \leq 2$ is always true. Furthermore, by the Hankel structure of the error matrix $\boldsymbol{E}$ and the choice of L, we have $\|\boldsymbol{E}\| \leq \sqrt{\min\{L, N-L+1\}}\left\|\boldsymbol{\mathcal{E}}-\widetilde{\boldsymbol{\mathcal{E}}}\right\| \leq \sqrt{L}\left\|\boldsymbol{\mathcal{E}}-\widetilde{\boldsymbol{\mathcal{E}}}\right\|$. Therefore, the assumptions of Theorem 4.3.14 are satisfied and we us the bound on $\mathrm{md}_{\mathbb{T}}(\Omega,\widetilde{\Omega})$ to obtain

$$\begin{aligned}\left\|\boldsymbol{A}_N - \widetilde{\boldsymbol{A}}_N\right\| \leq \left\|\boldsymbol{A}_N - \widetilde{\boldsymbol{A}}_N\right\|_{\mathrm{F}} &= \left(\sum_{j=1}^{M}\sum_{k=0}^{N-1}\left|z_j^k - \widetilde{z}_j^k\right|^2\right)^{\frac{1}{2}} \leq \sqrt{MN}2\pi N\left|t_j - \widetilde{t}_j\right|_{\mathbb{T}} \\ &\leq \sqrt{MN}2\pi N\,\mathrm{md}_{\mathbb{T}}(\Omega,\widetilde{\Omega}) \leq \sqrt{MN}2\pi N\frac{6(2M-1)}{L\cdot\alpha_{\min}}\left(\frac{qL+1}{qL-1}\right)^{\frac{5}{2}}\|\boldsymbol{E}\| \\ &\leq \frac{12\pi\sqrt{MN}N(2M-1)}{\sqrt{L}\cdot\alpha_{\min}}\left(\frac{qL+1}{qL-1}\right)^{\frac{5}{2}}\left\|\boldsymbol{\mathcal{E}}-\widetilde{\boldsymbol{\mathcal{E}}}\right\| \\ &\leq \frac{17\pi\sqrt{M}N(2M-1)}{\alpha_{\min}}\left(\frac{qL+1}{qL-1}\right)^{\frac{5}{2}}\left\|\boldsymbol{\mathcal{E}}-\widetilde{\boldsymbol{\mathcal{E}}}\right\|.\end{aligned} \tag{4.4.1}$$

Lemma 4.4.1 assumes $\left\|\boldsymbol{A}_N - \widetilde{\boldsymbol{A}}_N\right\| \leq \frac{\sigma_{\min}(\boldsymbol{A}_N)}{2}$, which is now fulfilled, using the lower bound

$$\sigma_{\min}(\boldsymbol{A}_N) \geq \sqrt{N}\sqrt{1-\frac{1}{qN}}$$

from Theorem 3.1.7, if

$$\frac{17\pi\sqrt{M}N(2M-1)}{\alpha_{\min}}\left(\frac{qL+1}{qL-1}\right)^{\frac{5}{2}}\left\|\boldsymbol{\mathcal{E}}-\widetilde{\boldsymbol{\mathcal{E}}}\right\| \leq \frac{1}{2}\sqrt{N}\sqrt{1-\frac{1}{qN}}.$$

This in turn is true by the assumption of this theorem.

Now we can apply Lemma 4.4.1 and Theorem 3.1.7 to obtain

$$\begin{aligned}
\frac{\|\boldsymbol{\alpha}-\widetilde{\boldsymbol{\alpha}}\|}{\|\boldsymbol{\alpha}\|} &\leq \sqrt{\frac{qN+1}{qN-1}}\left(\frac{1}{\|\boldsymbol{\mathcal{E}}\|}+\frac{17\pi\sqrt{M}N(2M-1)}{\alpha_{\min}\|\boldsymbol{A}_N\|}\left(\frac{qL+1}{qL-1}\right)^{\frac{5}{2}}\right)\left\|\boldsymbol{\mathcal{E}}-\widetilde{\boldsymbol{\mathcal{E}}}\right\| \\
&\leq \sqrt{3}\left(1+\frac{17\pi\sqrt{M}N(2M-1)\|\boldsymbol{\mathcal{E}}\|}{\alpha_{\min}\|\boldsymbol{A}_N\|}\left(\frac{qN+2}{qN-2}\right)^{\frac{5}{2}}\right)\frac{\left\|\boldsymbol{\mathcal{E}}-\widetilde{\boldsymbol{\mathcal{E}}}\right\|}{\|\boldsymbol{\mathcal{E}}\|} \\
&\leq \sqrt{3}\left(1+\frac{17\pi\sqrt{M}N(2M-1)\|\boldsymbol{A}_N\|\|\boldsymbol{\alpha}\|}{\alpha_{\min}\|\boldsymbol{A}_N\|}\left(\frac{qN+2}{qN-2}\right)^{\frac{5}{2}}\right)\frac{\left\|\boldsymbol{\mathcal{E}}-\widetilde{\boldsymbol{\mathcal{E}}}\right\|}{\|\boldsymbol{\mathcal{E}}\|} \\
&\leq \sqrt{3}\left(1+17\pi MN(2M-1)\frac{\alpha_{\max}}{\alpha_{\min}}\left(\frac{qN+2}{qN-2}\right)^{\frac{5}{2}}\right)\frac{\left\|\boldsymbol{\mathcal{E}}-\widetilde{\boldsymbol{\mathcal{E}}}\right\|}{\|\boldsymbol{\mathcal{E}}\|}.
\end{aligned}$$

□

Theorem 4.4.3 (Coefficient reconstruction error for clustered node configurations)**.** *Let $N, \lambda \in \mathbb{N}_+$, $N > 8\lambda$ and Ω be a clustered node configuration with respect to N and at most λ nodes in the biggest cluster. Let the minimum separation distance be q and assume the minimum cluster separation fulfills*

$$\Delta N \geq 8.8\lambda\left(\frac{4\mathrm{e}}{qN}\right)^{\frac{\lambda-1}{2}}.$$

If the error in the samples satisfies

$$\left\|\boldsymbol{\mathcal{E}}-\widetilde{\boldsymbol{\mathcal{E}}}\right\| \leq \frac{\alpha_{\min}}{2996\pi\lambda^{\frac{3}{2}}\sqrt{MN}(2M-1)}\left(\frac{qN}{2\mathrm{e}}\right)^{\lambda-1}\left(\frac{qN}{4\mathrm{e}}\right)^{5\lambda-5}\left(\frac{N-2}{N+1}\right)^{2\lambda-1},$$

then the relative error for the reconstruction of the exponential sum (4.1.1) *with the ESPRIT algorithm (paramter $L = \lceil \frac{N}{2} \rceil$) satisfies*

$$\begin{aligned}
\frac{\|\boldsymbol{\alpha}-\widetilde{\boldsymbol{\alpha}}\|}{\|\boldsymbol{\alpha}\|} &\leq 1.8\sqrt{2}\sqrt{\lambda}\left(\frac{2\mathrm{e}}{qN}\right)^{\lambda-1} \\
&\quad\left(1+1176\pi\lambda^{\frac{3}{2}}MN(2M-1)\frac{\alpha_{\max}}{\alpha_{\min}}\left(\frac{2\mathrm{e}}{qN}\right)^{5\lambda-5}\left(\frac{\lceil\frac{N}{2}\rceil}{\lceil\frac{N}{2}\rceil-1}\right)^{2\lambda-1}\right)\frac{\left\|\boldsymbol{\mathcal{E}}-\widetilde{\boldsymbol{\mathcal{E}}}\right\|}{\|\boldsymbol{\mathcal{E}}\|}.
\end{aligned}$$

Proof. First of all, since we are dealing with Vandermonde matrices $\boldsymbol{A}_L$ and $\boldsymbol{A}_N$, we verify that the node set Ω is a clustered node configuration such that Corollary 3.4.13 and Theorem 4.3.15 can be applied for both matrices. We have $L = \left\lceil \frac{N}{2} \right\rceil \geq \frac{N}{2}$ and hence $N \leq 2L$. Furthermore, it simply holds $L \leq N$. Since Ω is assumed to be a clustered node configuration with respect to N, all its clusters are contained in an interval of length $\frac{1}{N}$ and therefore, contained in intervals of length $\frac{1}{N} \leq \frac{1}{L}$ as well. The assumption on the cluster separation of the statement in this theorem are obviously stronger than Corollary 3.4.13 demands for being applied to $\boldsymbol{A}_N$. Moreover, $N \leq 2L$ yields together with the assumptions

$$2\Delta L \geq \Delta N \geq 8.8\lambda \left(\frac{4\mathrm{e}}{qN}\right)^{\frac{\lambda-1}{2}} \geq 8.8\lambda \left(\frac{2\mathrm{e}}{qL}\right)^{\frac{\lambda-1}{2}},$$

which is equivalent to

$$\Delta L \geq 4.4\lambda \left(\frac{2\mathrm{e}}{qL}\right)^{\frac{\lambda-1}{2}},$$

so that we can also apply Theorem 4.3.15.

Now we can proceed analogously to the proof of Theorem 4.4.2. By Theorem 4.3.15, we have

$$\begin{aligned}
\mathrm{md}_{\mathbb{T}}(\Omega, \widetilde{\Omega}) &\leq \frac{294\lambda^{\frac{3}{2}}(2M-1)}{L \cdot \alpha_{\min}} \left(\frac{2\mathrm{e}}{qL}\right)^{5\lambda-5} \left(\frac{L}{L-1}\right)^{2\lambda-1} \|\boldsymbol{E}\| \\
&\leq \frac{294\lambda^{\frac{3}{2}}(2M-1)}{\sqrt{L} \cdot \alpha_{\min}} \left(\frac{2\mathrm{e}}{qL}\right)^{5\lambda-5} \left(\frac{L}{L-1}\right)^{2\lambda-1} \left\|\boldsymbol{\mathcal{E}} - \widetilde{\boldsymbol{\mathcal{E}}}\right\|.
\end{aligned}$$

Using this in (4.4.1) yields

$$\begin{aligned}
\left\|\boldsymbol{A}_N - \widetilde{\boldsymbol{A}}_N\right\| &\leq \sqrt{MN} 2\pi N \, \mathrm{md}_{\mathbb{T}}(\Omega, \widetilde{\Omega}) \\
&\leq \sqrt{MN} 2\pi N \frac{294\lambda^{\frac{3}{2}}(2M-1)}{\sqrt{L} \cdot \alpha_{\min}} \left(\frac{2\mathrm{e}}{qL}\right)^{5\lambda-5} \left(\frac{L}{L-1}\right)^{2\lambda-1} \left\|\boldsymbol{\mathcal{E}} - \widetilde{\boldsymbol{\mathcal{E}}}\right\| \qquad (4.4.2) \\
&\leq \frac{832\pi\lambda^{\frac{3}{2}}\sqrt{M}N(2M-1)}{\alpha_{\min}} \left(\frac{2\mathrm{e}}{qL}\right)^{5\lambda-5} \left(\frac{L}{L-1}\right)^{2\lambda-1} \left\|\boldsymbol{\mathcal{E}} - \widetilde{\boldsymbol{\mathcal{E}}}\right\|
\end{aligned}$$

Additionally, Corollary 3.4.13 provides

$$\sigma_{\min}(\boldsymbol{A}_N) \geq \frac{\sqrt{N}}{1.8} \left(\frac{qN}{2\mathrm{e}}\right)^{\lambda-1}.$$

Using both we can check the assumption $\left\|\boldsymbol{A}_N - \widetilde{\boldsymbol{A}}_N\right\| \leq \frac{\sigma_{\min}(\boldsymbol{A}_N)}{2}$ of Lemma 4.4.1, which is therefore fulfilled if

$$\frac{832\pi\lambda^{\frac{3}{2}}\sqrt{M}N(2M-1)}{\alpha_{\min}} \left(\frac{2\mathrm{e}}{qL}\right)^{5\lambda-5} \left(\frac{L}{L-1}\right)^{2\lambda-1} \left\|\boldsymbol{\mathcal{E}} - \widetilde{\boldsymbol{\mathcal{E}}}\right\| \leq \frac{\sqrt{N}}{3.6} \left(\frac{qN}{2\mathrm{e}}\right)^{\lambda-1}.$$

This in turn holds as long as

$$\left\|\boldsymbol{\mathcal{E}} - \widetilde{\boldsymbol{\mathcal{E}}}\right\| \leq \frac{\alpha_{\min}}{2996\pi\lambda^{\frac{3}{2}}\sqrt{MN}(2M-1)} \left(\frac{qN}{2\mathrm{e}}\right)^{\lambda-1} \left(\frac{qL}{2\mathrm{e}}\right)^{5\lambda-5} \left(\frac{L-1}{L}\right)^{2\lambda-1},$$

which is true when the assumption is combined with

$$\frac{L-1}{L} = \frac{\left\lceil \frac{N}{2} \right\rceil - 1}{\left\lceil \frac{N}{2} \right\rceil} = \frac{2\left\lceil \frac{N}{2} \right\rceil - 2}{2\left\lceil \frac{N}{2} \right\rceil} \geq \frac{2\frac{N}{2} - 2}{N+1} = \frac{N-2}{N+1}. \tag{4.4.3}$$

which is true if the assumption hold. Now we can apply Lemma 4.4.1 combined with the bounds from Corollary 3.4.13 and Lemma 3.4.6, and (4.4.2) to obtain

$$\begin{aligned}\frac{\|\boldsymbol{\alpha} - \widetilde{\boldsymbol{\alpha}}\|}{\|\boldsymbol{\alpha}\|} \leq\ & 1.8\sqrt{2}\sqrt{\lambda}\left(\frac{2\mathrm{e}}{qN}\right)^{\lambda-1} \\ & \left(1 + 832\pi\lambda^{\frac{3}{2}}\sqrt{MN}(2M-1)\frac{\alpha_{\max}}{\alpha_{\min}}\left(\frac{2\mathrm{e}}{qN}\right)^{5\lambda-5}\left(\frac{L}{L-1}\right)^{2\lambda-1}\right)\frac{\left\|\boldsymbol{\mathcal{E}} - \widetilde{\boldsymbol{\mathcal{E}}}\right\|}{\|\boldsymbol{\mathcal{E}}\|}\end{aligned}$$

Finally, using (4.4.3) yields the result. □

Appendix A

Schur-complement technique with Fejér kernel

Here we present the Schur-complement technique for pair clusters with applied to a Fejér kernel matrix in order to drop the dependencies on the number of nodes M in the results from Section 3.3. Results from here are referred to in Section 3.3.4 and we use the same notation. The calculations are completely analogous to the ones done for the Dirichlet kernel matrices in Section 3.3.

The Fejér kernel of degree $n \in \mathbb{N}$ is given by $F_n \colon \mathbb{R} \to \mathbb{R}$,

$$F_n(t) := \sum_{k=-n}^{n} \left(1 - \frac{|k|}{n+1}\right) \mathrm{e}^{2\pi i k t} = \begin{cases} n+1, & t \in \mathbb{Z}, \\ \frac{1}{n+1} \left(\frac{\sin((n+1)\pi t)}{\sin(\pi t)}\right)^2, & \text{otherwise}, \end{cases} \tag{A.0.1}$$

so that for $\boldsymbol{C} := \operatorname{diag}(1 - |k|/(n+1))_{|k| \le n} \in \mathbb{R}^{N \times N}$ we define

$$\widetilde{\boldsymbol{K}} := \operatorname{diag}(z_1^{-n}, \ldots, z_M^{-n}) \boldsymbol{A} \boldsymbol{C} \boldsymbol{A}^* \operatorname{diag}(z_1^{-n}, \ldots, z_M^{-n})^* = \left(F_n(t_i - t_j)\right)_{i,j=1}^{M} \in \mathbb{R}^{M \times M}. \tag{A.0.2}$$

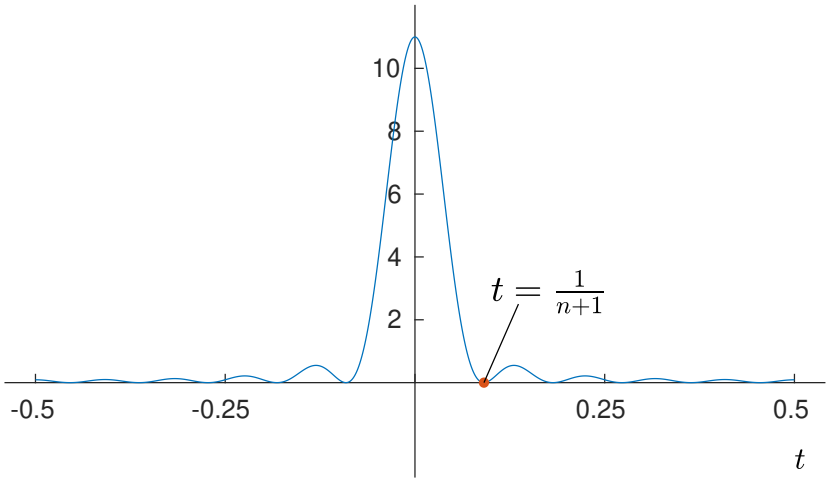


Figure A.0.1: Fejér kernel of degree $n = 10$ over the interval $[-1/2, 1/2]$

Analogously, to the bounds for the Dirichlet kernel we have the following bounds for the Fejér kernel. We notice that the Fejér kernel has quadratic decay.

Lemma A.0.1 (Fejér kernel bounds, cf. Lemma 2.4.7)**.**
Let $n \in \mathbb{N}$. Then the Fejér kernel (A.0.1) *is bounded by*

$$(n+1)\left(1-\frac{\pi^2}{6}(n+1)^2t^2\right)^2 \leq F_n(t) \leq (n+1)\left(1-(n+1)^2t^2\right)^2, \qquad 0 \leq |t| \leq \frac{1}{2n+1}.$$

Furthermore, the Fejér kernel and its first two derivatives are bounded by

$$\begin{aligned}
|F_n(t)| &\leq (n+1)\frac{1}{4(n+1)^2|t|^2},\\
\left|F_n'(t)\right| &\leq (n+1)^2\left(\frac{\pi}{2(n+1)^2|t|^2}+\frac{1}{2(n+1)^3|t|^3}\right),\\
\left|F_n''(t)\right| &\leq (n+1)^3\left(\frac{\pi^2}{(n+1)^2|t|^2}+\frac{2\pi}{(n+1)^3|t|^3}+\frac{8+\pi^2}{8(n+1)^4|t|^4}\right)
\end{aligned}$$

for $0 < |t| \leq 1/2$.

Proof. Due to symmetry, it suffices to prove all bounds for $t > 0$ and we use the explicit expression of the Fejér kernel on the right hand side in (A.0.1). The lower bound on $F_n(t)$ can be derived from the inequalities $x - x^3/6 \leq \sin(x) \leq x$, that hold for all $x \in [0,\pi]$. The left inequality with $x = (n+1)\pi t$ and the right inequality with $x = \pi t$ lead to

$$\sin((n+1)\pi t) \geq \left((n+1)-\frac{\pi^2}{6}(n+1)^3t^2\right)\pi t \geq \left((n+1)-\frac{\pi^2}{6}(n+1)^3t^2\right)\sin(\pi t).$$

Note that $1-\frac{\pi^2}{6}(n+1)^2t^2 > 0$ for $t \leq \frac{1}{2n+1}$ and therefore, the inequality holds after squaring. The upper bound on $F_n(t)$ can be derived from the inequality $\cos(\alpha x) \leq \cos(x)$ that holds for all $x \in [0,\pi/2]$ and $\alpha > 1$ such that $\alpha x \in [0,\pi/2]$. Integrating this inequality, choosing $\alpha = (n+1)/2$ and $x = \pi t$, and applying the double angle formula yields

$$\frac{\sin((n+1)\pi t)}{2\cos(\frac{n+1}{2}\pi t)} = \sin\left(\frac{n+1}{2}\pi t\right) \leq \frac{n+1}{2}\sin(\pi t).$$

Reordering the inequality and applying that $\cos(x) \leq 1 - 4x^2/\pi^2$ for all $x \in [0,\pi/2]$ yields

$$\frac{\sin((n+1)\pi t)}{\sin(\pi t)} \leq (n+1)\cos(\frac{n+1}{2}\pi t) \leq (n+1)(1-(n+1)^2t^2).$$

Finally, the remaining bounds on the absolute values can be proven by calculating the first and second derivatives and using $\sin(x) \geq 2x/\pi$ and $\cot(x) \leq 1/x$ that hold for all $x \in (0,\pi/2]$. The fist derivative is given by

$$F_n'(t) = \frac{2\pi}{n+1}\frac{\sin(\pi(n+1)t)}{\sin^2(\pi t)}\left((n+1)\cos(\pi(n+1)t) - \cot(\pi t)\sin(\pi(n+1)t)\right)$$

and the second derivative is given by

$$F_n''(t) = \frac{2\pi^2(n+1)}{\sin^2(\pi t)}\left(\cos^2(\pi(n+1)t) - \sin^2(\pi(n+1)t)\right)$$

$$+\frac{2\pi^2}{n+1}\sin^2(\pi(n+1)t)\left(\frac{1}{\sin^4(\pi t)}+2\frac{\cot^2(\pi t)}{\sin^2(\pi t)}\right)$$
$$-8\pi^2\cot(\pi t)\frac{\sin(\pi(n+1)t)}{\sin^2(\pi t)}\cos(\pi(n+1)t)$$

for $t>0$.

□

Lemma A.0.2 (Extremal eigenvalues of Fejér kernel matrix, cf. [71, Thm. 4.1])**.**
The extremal eigenvalues of the Hermitian, positive semidefinite Fejér kernel matrix $\widetilde{\boldsymbol{K}}$ are bounded by

$$(n+1)\left(1-\frac{\pi^2}{3(qN)^2}\right)\leq\lambda_{\min}(\widetilde{\boldsymbol{K}})\leq n+1\leq\lambda_{\max}(\widetilde{\boldsymbol{K}})\leq(n+1)\left(1+\frac{\pi^2}{3(qN)^2}\right).$$

Particularly, $\left\|\widetilde{\boldsymbol{K}}^{-1}\right\|\leq\frac{1}{n+1}\cdot\frac{3(qN)^2}{3(qN)^2-\pi^2}$ and if the corresponding nodes are $q>\frac{\pi}{\sqrt{3}N}$ separated, then the lower bound on $\lambda_{\min}(\widetilde{\boldsymbol{K}})$ is non-trivial, $\widetilde{\boldsymbol{K}}$ has full rank and is positive definite.

Proof. We follow [71] and use Gerschgorin's circle theorem. For each eigenvalue λ of $\widetilde{\boldsymbol{K}}$ we have, using Lemma A.0.1 and the monotony of the upper bound on the Fejér kernel,

$$|n+1-\lambda|\leq\frac{2}{n+1}\sum_{j=1}^{\lceil\frac{M}{2}\rceil}\frac{1}{4qj^2}\leq\frac{1}{2(n+1)}\sum_{j=1}^{\infty}\frac{1}{j^2}=(n+1)\frac{\pi^2}{12q^2(n+1)^2}$$
$$=(n+1)\frac{\pi^2}{3q^2(2n+2)^2}\leq(n+1)\frac{\pi^2}{3(qN)^2}.$$

□

Applying this for the extremal eigenvalues yields

$$\lambda_{\min}(\widetilde{\boldsymbol{K}})\geq n+1-(n+1)\frac{\pi^2}{3(qN)^2}=(n+1)\left(1-\frac{\pi^2}{3(qN)^2}\right)\geq\frac{N}{2}\left(1-\frac{\pi^2}{3(qN)^2}\right)$$

and

$$\lambda_{\max}(\widetilde{\boldsymbol{K}})\leq n+1+(n+1)\frac{\pi^2}{3(qN)^2}=(n+1)\left(1+\frac{\pi^2}{3(qN)^2}\right)\leq N\left(1-\frac{\pi^2}{3(qN)^2}\right).$$

Nodes with one pair cluster

Definition A.0.3 (cf. Definition 3.3.2)**.**
For $N=2n+1\in\mathbb{N}_+$ we define

$$\boldsymbol{a}_1:=\left(z_1^k\right)_{k\in\mathbb{Z}\,|k|\leq n}\in\mathbb{C}^{1\times N}\quad\text{and}\quad\boldsymbol{A}_2:=\left(z_j^k\right)_{\substack{j=2,\dots,M\\k\in\mathbb{Z},\,|k|\leq n}}\in\mathbb{C}^{(M-1)\times N}$$

so that with $\widetilde{\boldsymbol{K}}$ from (A.0.2) *and*

$$\boldsymbol{a}_1\boldsymbol{C}\boldsymbol{a}_1^*=n+1,\quad\widetilde{\boldsymbol{K}}_2:=\boldsymbol{A}_2C\boldsymbol{A}_2^*\quad\text{and}\quad\boldsymbol{b}:=\boldsymbol{A}_2\boldsymbol{C}\boldsymbol{a}_1^*=\begin{pmatrix}F_n(q)\\F_n(t_3)\\\vdots\\F_n(t_M)\end{pmatrix},\tag{A.0.3}$$

we have the partitioning

$$\boldsymbol{A} = \begin{pmatrix} \boldsymbol{a}_1 \\ \boldsymbol{A}_2 \end{pmatrix} \quad \textit{and} \quad \widetilde{\boldsymbol{K}} = \begin{pmatrix} n+1 & \boldsymbol{b}^* \\ \boldsymbol{b} & \boldsymbol{K}_2 \end{pmatrix}. \tag{A.0.4}$$

We note that under the assumption from (3.3.3) *the shifted Vandermonde matrix* $\boldsymbol{A}_2$ *has nodes that are at least* Δ *separated.*

Lemma A.0.4 (cf. Lemma 3.3.4)**.**
Under the conditions of (3.3.3) *and with* $\boldsymbol{b}$ *as in* (A.0.3)*, we have*

$$\boldsymbol{b} = \boldsymbol{K}_2 \boldsymbol{e}_1 + \boldsymbol{r},$$

where $\boldsymbol{e}_1 \in \mathbb{R}^{(M-1)}$ *denotes the first unit vector and* $\boldsymbol{r} = (r_1, \ldots, r_{M-1})^\top \in \mathbb{R}^{M-1}$ *satisfies*

$$\|\boldsymbol{r}\|^2 \leq (n+1-F_n(q))^2 + (n+1)^4 q^2 \left(\frac{\pi^6}{180((n+1)\Delta)^4} + \frac{1.04\pi}{((n+1)\Delta)^5} + \frac{\pi^6}{1890((n+1)\Delta)^6} \right).$$

Proof. The vector $\boldsymbol{b}$ can be approximated by the first column of $\boldsymbol{K}_2$ in the sense that

$$\boldsymbol{b} = \begin{pmatrix} F_n(q) \\ F_n(t_3) \\ \vdots \\ F_n(t_M) \end{pmatrix} = \begin{pmatrix} F_n(0) \\ F_n(t_3 - q) \\ \vdots \\ F_n(t_M - q) \end{pmatrix} + \begin{pmatrix} \hat{r}_1 \\ \vdots \\ \hat{r}_{M-1} \end{pmatrix}.$$

We have $|\hat{r}_1| = n + 1 - F_n(q)$ and for $j = 2, \ldots, M-1$ the mean value theorem yields

$$|\hat{r}_j| = |F_n(t_{j+1}) - F_n(t_{j+1} - q)| = \left|F_n'(\xi_j)\right| q, \quad \xi_j \in (|t_{j+1} - q|_{\mathbb{T}}, |t_{j+1}|_{\mathbb{T}}).$$

Note that, in the worst case, half of the nodes can be as close as possible (under the assumed separation condition) to t_2 not only on its right but also on its left. Hence, for $j = 2, \ldots, \lceil \frac{M}{2} \rceil$, $\xi_j \geq (j-1)\Delta$ and Lemma A.0.1 lead to

$$\begin{aligned} |\hat{r}_j| &\leq (n+1)^2 \left(\frac{\pi}{2(n+1)^2 |\xi_j|^2} + \frac{1}{2(n+1)^3 |\xi_j|^3} \right) q \\ &\leq (n+1)^2 \left(\frac{\pi}{2(j-1)^2((n+1)\Delta)^2} + \frac{1}{2(j-1)^3((n+1)\Delta)^3} \right) q. \end{aligned}$$

Thus, for all nodes we get

$$\begin{aligned} \sum_{j=2}^{M-1} |\hat{r}_j|^2 &\leq 2(n+1)^4 q^2 \sum_{j=2}^{\lceil M/2 \rceil} \left(\frac{\pi^2}{4((n+1)\Delta)^4 j^4} + \frac{\pi}{2((n+1)\Delta)^5 j^5} + \frac{1}{4((n+1)\Delta)^6 j^6} \right) \\ &\leq (n+1)^4 q^2 \Bigg(\frac{\pi^2}{2((n+1)\Delta)^4} \underbrace{\sum_{j=1}^{\infty} \frac{1}{j^4}}_{=\frac{\pi^4}{90}} + \frac{\pi}{((n+1)\Delta)^5} \underbrace{\sum_{j=1}^{\infty} \frac{1}{j^5}}_{\leq 1.04} + \frac{1}{2((n+1)\Delta)^6} \underbrace{\sum_{j=1}^{\infty} \frac{1}{j^6}}_{=\frac{\pi^6}{945}} \Bigg). \end{aligned}$$

□

Lemma A.0.5 (cf. Lemma 3.3.5)**.**
Under the conditions of (3.3.3) *and Definition A.0.3, if* $\Delta N \geq 5$*, we have*

$$\left\|\widetilde{\boldsymbol{K}}^{-1}\right\| \leq \frac{C(\Delta N)}{N(qN)^2},$$

where

$$C(\Delta N) = 8\left(\frac{6(\Delta N)^2-\pi^2}{3(\Delta N)^2-\pi^2} + \sqrt{\frac{3(\Delta N)^2}{3(\Delta N)^2-\pi^2}}\right) \cdot \left[2 - \frac{3(\Delta N)^2}{3(\Delta N)^2-\pi^2}\left(1 + \frac{4\pi^2}{45(\Delta N)^4} + \frac{1.04\cdot 32\pi}{(\Delta N)^5} + \frac{32\pi^6}{945(\Delta N)^6}\right)\right]^{-1}.$$

Proof. We consider $\widetilde{\boldsymbol{K}}$ decomposed as in (3.3.5) and apply Lemma 2.1.21 with respect to $\boldsymbol{K}_2$ to obtain

$$\widetilde{\boldsymbol{K}}^{-1} = \begin{pmatrix} 1 & \boldsymbol{0}_{1\times(M-1)} \\ -\boldsymbol{K}_2^{-1}\boldsymbol{b} & \boldsymbol{I}_{M-1}\end{pmatrix}\begin{pmatrix}(n+1-\boldsymbol{b}^*\boldsymbol{K}_2^{-1}\boldsymbol{b})^{-1} & \boldsymbol{0}_{1\times(M-1)} \\ \boldsymbol{0}_{(M-1)\times 1} & \boldsymbol{K}_2^{-1}\end{pmatrix}\begin{pmatrix}1 & -\boldsymbol{b}^*\boldsymbol{K}_2^{-1} \\ \boldsymbol{0}_{(M-1)\times 1} & \boldsymbol{I}_{M-1}\end{pmatrix}$$

and thus,

$$\left\|\widetilde{\boldsymbol{K}}^{-1}\right\| \leq \left\|\begin{pmatrix} 1 & \boldsymbol{0}_{1\times(M-1)} \\ -\boldsymbol{K}_2^{-1}\boldsymbol{b} & \boldsymbol{I}_{M-1}\end{pmatrix}\right\|^2 \max\left\{\left\|\boldsymbol{K}_2^{-1}\right\|, \left|\left(n+1-\boldsymbol{b}^*\boldsymbol{K}_2^{-1}\boldsymbol{b}\right)^{-1}\right|\right\}.$$

First of all, we establish an upper bound for the norm of the triangular matrix. By singular value decomposition we have $\left((\boldsymbol{A}_2\boldsymbol{C}\boldsymbol{A}_2^*)^{-1}\boldsymbol{A}_2\boldsymbol{C}^{\frac{1}{2}}\right)^* = (\boldsymbol{A}_2\boldsymbol{C}^{\frac{1}{2}})^\dagger$ and (A.0.3) and Lemma A.0.2 imply

$$\begin{aligned}\left\|\boldsymbol{K}_2^{-1}\boldsymbol{b}\right\| = \left\|(\boldsymbol{A}_2\boldsymbol{C}\boldsymbol{A}_2^*)^{-1}\boldsymbol{A}_2\boldsymbol{C}a_1^*\right\| &\leq \sqrt{\left\|\boldsymbol{K}_2^{-1}\right\|}\cdot\sqrt{\|a_1\boldsymbol{C}a_1^*\|} \\ &= \sqrt{\frac{3(\Delta N)^2}{3(\Delta N)^2-\pi^2}}\cdot\sqrt{F_n(0)} = \sqrt{\frac{3(\Delta N)^2}{3(\Delta N)^2-\pi^2}}.\end{aligned}$$

Together with Lemma 2.1.22, we obtain

$$\left\|\begin{pmatrix} 1 & \boldsymbol{0}_{1\times(M-1)} \\ -\boldsymbol{K}_2^{-1}b & \boldsymbol{I}_{M-1}\end{pmatrix}\right\|^2 \leq 1 + \left\|\boldsymbol{K}_2^{-1}\boldsymbol{b}\right\| + \left\|\boldsymbol{K}_2^{-1}\boldsymbol{b}\right\|^2 \leq \frac{6(\Delta N)^2-\pi^2}{3(\Delta N)^2-\pi^2} + \sqrt{\frac{3(\Delta N)^2}{3(\Delta N)^2-\pi^2}}. \tag{A.0.5}$$

The next step is to bound $(n+1-\boldsymbol{b}^*\boldsymbol{K}_2^{-1}\boldsymbol{b})^{-1}$. Lemma A.0.4 yields

$$\boldsymbol{b}^*\boldsymbol{K}_2^{-1}\boldsymbol{b} = (\boldsymbol{K}_2e_1+\boldsymbol{r})^*\boldsymbol{K}_2^{-1}(\boldsymbol{K}_2e_1+\boldsymbol{r}) = 2F_n(q) - (n+1) + \boldsymbol{r}^*\boldsymbol{K}_2^{-1}\boldsymbol{r}.$$

Applying the second part of Lemma A.0.4, Lemma A.0.1, and Lemma A.0.2 yields

$$\begin{aligned} n+1-\boldsymbol{b}^*\boldsymbol{K}_2^{-1}\boldsymbol{b} &\geq 2\left(n+1-F_n(q)\right) - \|\hat{\boldsymbol{r}}\|^2\left\|\boldsymbol{K}_2^{-1}\right\| \\ &\geq \left(n+1-F_n(q)\right)\left(2-\left(n+1-F_n(q)\right)\left\|\boldsymbol{K}_2^{-1}\right\|\right)\end{aligned}$$

$$
\begin{aligned}
&\quad - \left\|\boldsymbol{K}_2^{-1}\right\| (n+1)^4 q^2 \left(\frac{\pi^6}{180((n+1)\Delta)^4} + \frac{1.04\pi}{((n+1)\Delta)^5} + \frac{\pi^6}{1890((n+1)\Delta)^6} \right) \\
&\geq (n+1) \left(2((n+1)q)^2 - ((n+1)q)^4\right) \left(2 - (n+1)\left\|\boldsymbol{K}_2^{-1}\right\|\right) \\
&\quad - \left\|\boldsymbol{K}_2^{-1}\right\| (n+1)^4 q^2 \left(\frac{\pi^6}{180((n+1)\Delta)^4} + \frac{1.04\pi}{((n+1)\Delta)^5} + \frac{\pi^6}{1890((n+1)\Delta)^6} \right) \\
&\geq (n+1)^3 q^2 \Bigg[\left(2 - ((n+1)q)^2\right) \left(2 - (n+1)\left\|\boldsymbol{K}_2^{-1}\right\|\right) \\
&\quad - (n+1)\left\|\boldsymbol{K}_2^{-1}\right\| \left(\frac{4\pi^6}{45((2n+2)\Delta)^4} + \frac{1.04 \cdot 32\pi}{((2n+2)\Delta)^5} + \frac{32\pi^6}{945((2n+2)\Delta)^6} \right) \Bigg] \\
&\geq (n+1)^3 q^2 \Bigg[\left(2 - (n+1)\left\|\boldsymbol{K}_2^{-1}\right\|\right) \\
&\quad - (n+1)\left\|\boldsymbol{K}_2^{-1}\right\| \left(\frac{4\pi^6}{45((2n+2)\Delta)^4} + \frac{1.04 \cdot 32\pi}{((2n+2)\Delta)^5} + \frac{32\pi^6}{945((2n+2)\Delta)^6} \right) \Bigg] \\
&\geq \frac{N(qN)^2}{8} \left[2 - \frac{3(\Delta N)^2}{3(\Delta N)^2 - \pi^2} \left(1 + \frac{4\pi^6}{45(\Delta N)^4} + \frac{1.04 \cdot 32\pi}{(\Delta N)^5} + \frac{32\pi^6}{945(\Delta N)^6} \right) \right].
\end{aligned}
$$

For $\Delta N \geq 5$, the most inner bracketed term takes values in $(0.5, 1.4)$ such that the square bracketed term is positive. Forming the reciprocal gives the result, since Lemma A.0.2 also implies

$$
N\left\|\boldsymbol{K}_2^{-1}\right\| \leq 8 \left[2 - \frac{3(\Delta N)^2}{3(\Delta N)^2 - \pi^2}(1 + \cdots) \right]^{-1}. \tag{A.0.6}
$$

□

Theorem A.0.6 (Upper bound, cf. Theorem 3.3.6)**.**
Under the conditions of (3.3.3) *with* $\Delta N \geq \Delta_{\min} N = 6$*, we have*

$$
\operatorname{cond}(\boldsymbol{A}) \leq \frac{8.2}{qN}.
$$

Proof. The bound follows from Lemma 3.3.3 and Lemma 2.1.17 together with Lemma A.0.5 and $C(\Delta N) \leq C(6) \leq 29$. □

Several pair clusters

In this section let Ω be a set of several pair clusters as in (3.3.9).

Definition A.0.7.
For $N = 2n+1 \in \mathbb{N}_+$ *and* $\widetilde{\boldsymbol{K}}$ *from* (A.0.2) *we define*

$$
\boldsymbol{A}_1 := \left(z_j^k\right)_{\substack{j=1,\dots,M/2 \\ k\in\mathbb{Z},\,|k|\leq n}} \in \mathbb{C}^{(M/2)\times N} \quad \text{and} \quad \boldsymbol{A}_2 := \left(z_j^k\right)_{\substack{j=M/2+1,\dots,M \\ k\in\mathbb{Z},\,|k|\leq n}} \in \mathbb{C}^{(M/2)\times N}
$$

so that with $\boldsymbol{K}_1 := \boldsymbol{A}_1 \boldsymbol{C} \boldsymbol{A}_1^*$, $\boldsymbol{K}_2 := \boldsymbol{A}_2 \boldsymbol{C} \boldsymbol{A}_2^*$ *and* $\boldsymbol{B} := \boldsymbol{A}_2 \boldsymbol{C} \boldsymbol{A}_1^*$ *all in* $\mathbb{C}^{(M/2)\times(M/2)}$ *we have the partitioning*

$$
\operatorname{diag}(z_1^{-n}, \dots, z_M^{-n})\boldsymbol{A} = \begin{pmatrix} \boldsymbol{A}_1 \\ \boldsymbol{A}_2 \end{pmatrix}, \quad \widetilde{\boldsymbol{K}} = \begin{pmatrix} \boldsymbol{K}_1 & \boldsymbol{B}^* \\ \boldsymbol{B} & \boldsymbol{K}_2 \end{pmatrix}. \tag{A.0.7}
$$

Note that under the assumptions in (3.3.9) *the shifted Vandermonde matrices* $\boldsymbol{A}_1$ *and* $\boldsymbol{A}_2$ *are each corresponding to nodes that are at least* Δ *separated.*

Lemma A.0.8.
Under the conditions of (3.3.9), $\boldsymbol{R}_1 := \boldsymbol{B} - \boldsymbol{K}_1$ *fulfills*

$$\|\boldsymbol{R}_1\| \leq n + 1 - F_n(cq) + c(n+1)^2 q \left(\frac{\pi^3}{6((n+1)\Delta)^2} + \frac{1.21}{((n+1)\Delta)^3} \right).$$

Proof. The Fejér kernel F_n is monotone decreasing on $[0, 1/N]$. Hence, for the diagonal entries we obtain

$$|(\boldsymbol{R}_1)_{jj}| = \left| F_n\left(t_j - t_{j+\frac{M}{2}}\right) - (n+1) \right| = n + 1 - F_n\left(t_j - t_{j+\frac{M}{2}}\right) \leq n + 1 - F_n(cq).$$

The off diagonal entries are bounded by the mean value theorem and Lemma A.0.1 as

$$\begin{aligned} |(\boldsymbol{R}_1)_{j\ell}| &= \left| F_n(t_j - t_\ell) - F_n\left(t_{j+\frac{M}{2}} - t_\ell\right) \right| \\ &\leq \left| F_n'(\xi_{j\ell}) \right| cq \leq c(n+1)^2 q \left(\frac{\pi}{2(n+1)^2 \xi_{j\ell}^2} + \frac{1}{2(n+1)^3 \xi_{j\ell}^3} \right), \end{aligned}$$

where $\left(\left| t_{j+\frac{M}{2}} - t_\ell \right|_{\mathbb{T}}, |t_j - t_\ell|_{\mathbb{T}} \right) \ni \xi_{j\ell} \geq |j - \ell|' \Delta$ implies

$$|(\boldsymbol{R}_1)_{j\ell}| \leq c(n+1)^2 q \left(\frac{\pi}{2((n+1)\Delta)^2 (|j-\ell|')^2} + \frac{1}{2((n+1)\Delta)^3 (|j-\ell|')^3} \right) =: (\tilde{\boldsymbol{R}}_1)_{j\ell}$$

for $j, \ell = 1, \ldots, \frac{M}{2}$, $j \neq \ell$. Additionally, we set $(\widetilde{\boldsymbol{R}}_1)_{jj} := n + 1 - F_n(cq)$. We bound the spectral norm of $\boldsymbol{R}_1$ by the one of the real symmetric matrix $\widetilde{\boldsymbol{R}}_1$ using Lemma 2.1.23 and proceed by

$$\begin{aligned} \|\boldsymbol{R}_1\| &\leq \left\| \widetilde{\boldsymbol{R}}_1 \right\| \leq \left\| \widetilde{\boldsymbol{R}}_1 \right\|_\infty \\ &\leq n + 1 - F_n(cq) + 2c(n+1)^2 q \sum_{j=1}^{\lfloor \frac{M}{4} \rfloor} \left(\frac{\pi}{2((n+1)\Delta)^2 j^2} + \frac{1}{2((n+1)\Delta)^3 j^3} \right), \end{aligned}$$

from which the assertion follows. □

Lemma A.0.9.
Under the conditions of (3.3.9), $\boldsymbol{R}_1 = \boldsymbol{B} - \boldsymbol{K}_1$ *and* $\boldsymbol{R}_2 := \boldsymbol{B} - \boldsymbol{K}_2$ *fulfill*

$$\begin{aligned} \left\| 2N \boldsymbol{I}_{M/2} + \boldsymbol{R}_1^* + \boldsymbol{R}_2 \right\| &\leq 2(n + F_n(q)) \\ &\quad + c^2 (n+1)^3 q^2 \left(\frac{\pi^4}{3((n+1)\Delta)^2} + \frac{4.84\pi}{((n+1)\Delta)^3} + \frac{8\pi^4 + \pi^6}{360((n+1)\Delta)^4} \right). \end{aligned}$$

Proof. First, note that

$$(\boldsymbol{R}_1^* + \boldsymbol{R}_2)_{j\ell} = F_n\left(t_{j+\frac{M}{2}} - t_\ell\right) + F_n\left(t_j - t_{\ell+\frac{M}{2}}\right) - F_n\left(t_{j+\frac{M}{2}} - t_{\ell+\frac{M}{2}}\right) - F_n(t_j - t_\ell).$$

Monotonicity of the Fejér kernel F_n on $t \in [0, 1/N]$ gives, for $j = \ell$,

$$\begin{aligned}
&\left|(2N\boldsymbol{I}_{M/2} + \boldsymbol{R}_1^* + \boldsymbol{R}_2)_{jj}\right| \\
&= 2N + \left(F_n\left(t_j - t_{j+\frac{M}{2}}\right) - (n+1)\right) + \left(F_n\left(t_{j+\frac{M}{2}} - t_j\right) - (n+1)\right) \\
&\leq 2(N - (n+1)) + 2F_n(q) \leq 2\left(n + F_n(q)\right).
\end{aligned}$$

For each fixed off diagonal entry $j \neq \ell$, the matrix $2N\boldsymbol{I}$ has no contribution. We write the node $t_{j+M/2}$ as a perturbation of t_j by $h_j := t_{j+M/2} - t_j$ and expand the Fejér kernel by its Taylor polynomial of degree 2 in the point $\hat{h} := t_j - t_\ell + \frac{h_j - h_\ell}{2}$. Using

$$F_n(h) = F_n\left(\hat{h}\right) + F_n'\left(\hat{h}\right)(h - \hat{h}) + \frac{F_n''(\xi)}{2}(h - \hat{h})^2$$

for some $\xi \in [\hat{h}, h] \cup [h, \hat{h}]$, the constant term as well as the linear term cancel out and we get

$$\begin{aligned}
&F_n(t_j + h_j - t_\ell) + F_n(t_j - t_\ell - h_\ell) - F_n(t_j + h_j - t_\ell - h_\ell) - F_n(t_j - t_\ell) \\
&= \frac{1}{8}\left(F_n''(\xi_1)(h_j + h_\ell)^2 + F_n''(\xi_2)(h_j + h_\ell)^2 + F_n''(\xi_3)(h_j - h_\ell)^2 + F_n''(\xi_4)(h_j - h_\ell)^2\right).
\end{aligned}$$

Lemma A.0.1 and $\xi_1, \ldots, \xi_4 \geq |j - \ell|' \Delta$ imply

$$\begin{aligned}
|(\boldsymbol{R}_1^* + \boldsymbol{R}_2)_{j\ell}| \leq \frac{1}{4}(n+1)^3 \Bigg(&\frac{\pi^2}{((n+1)\Delta)^2(|j-\ell|')^2} + \frac{2\pi}{((n+1)\Delta)^3(|j-\ell|')^3} \\
&+ \frac{8 + \pi^2}{8((n+1)\Delta)^4(|j-\ell|')^4}\Bigg) \cdot \left((h_j + h_\ell)^2 + (h_j - h_\ell)^2\right)
\end{aligned}$$

and hence by $h_j, h_\ell \leq cq$

$$\begin{aligned}
\left|(2N\boldsymbol{I}_{M/2} + \boldsymbol{R}_1^* + \boldsymbol{R}_2)_{j\ell}\right| \leq c^2(n+1)^3 q^2 \Bigg(&\frac{\pi^2}{((n+1)\Delta)^2(|j-\ell|')^2} + \frac{2\pi}{((n+1)\Delta)^3(|j-\ell|')^3} \\
&+ \frac{8 + \pi^2}{8((n+1)\Delta)^4(|j-\ell|')^4}\Bigg).
\end{aligned}$$

The matrix $2N\boldsymbol{I}_{M/2} + \boldsymbol{R}_1^* + \boldsymbol{R}_2$ is real symmetric so that

$$\begin{aligned}
\left\|2N\boldsymbol{I}_{M/2} + \boldsymbol{R}_1^* + \boldsymbol{R}_2\right\| \leq \left\|2N\boldsymbol{I}_{M/2} + \boldsymbol{R}_1^* + \boldsymbol{R}_2\right\|_\infty &\leq 2\left(n + F_n(q)\right) + 2c^2(n+1)^3 q^2 \\
&\cdot \sum_{j=1}^{\lfloor \frac{M}{4} \rfloor} \left(\frac{\pi^2}{((n+1)\Delta)^2 j^2} + \frac{2\pi}{((n+1)\Delta)^3 j^3} + \frac{8+\pi^2}{8((n+1)\Delta)^4 j^4}\right) \\
&\leq 2\left(n + F_n(q)\right) + 2c^2(n+1)^3 q^2 \\
&\cdot \left(\frac{\pi^4}{6((n+1)\Delta)^2} + \frac{1.21 \cdot 2\pi}{((n+1)\Delta)^3} + \frac{8\pi^4 + \pi^6}{720((n+1)\Delta)^4}\right)
\end{aligned}$$

and therefore the result holds. □

Lemma A.0.10.
Under the conditions of (3.3.9) *with* $N = 2n+1 \in \mathbb{N}_+$, $qN \leq 1/2$ *and* $\Delta N \geq 2$, *such that*

$$\widetilde{C}(qN, \Delta N, c) := \frac{1}{8}\Bigg(2 - \frac{4c^2\pi^4}{3(\Delta N)^2} - \frac{39c^2\pi}{(\Delta N)^3} - \frac{16c^2\pi^4 + 2c^2\pi^6}{45(\Delta N)^4} \\ - \frac{3(\Delta N)^2}{3(\Delta N)^2 - \pi^2}\left(\frac{2c^2\pi^2}{9}qN + \frac{2c\pi^3}{3(\Delta N)^2} + \frac{9.68c}{(\Delta N)^3}\right)^2\Bigg)$$

is positive, we have

$$\left\|K^{-1}\right\| \leq \frac{C(qN, \Delta N, c)}{N(qN)^2},$$

where

$$C(qN, \Delta N, c) := \left(\frac{6(\Delta N)^2}{3(\Delta N)^2 - \pi^2} + \sqrt{\frac{3(\Delta N)^2 + \pi^2}{3(\Delta N)^2 - \pi^2}}\right)/\widetilde{C}(qN, \Delta N, c).$$

Proof. We proceed analogously to Lemma A.0.5 and apply Lemma 2.1.21 to the matrix $\widetilde{\boldsymbol{K}}$ decomposed as in (A.0.7) and obtain

$$\|\widetilde{\boldsymbol{K}}^{-1}\| \leq \max\left\{\|\boldsymbol{K}_1^{-1}\|, \|(\boldsymbol{K}_2 - \boldsymbol{B}\boldsymbol{K}_1^{-1}\boldsymbol{B}^*)^{-1}\|\right\}\left\|\begin{pmatrix}\boldsymbol{I}_{M/2} & \boldsymbol{0}_{M/2}\\ -\boldsymbol{B}\boldsymbol{K}_1^{-1} & \boldsymbol{I}_{M/2}\end{pmatrix}\right\|^2. \tag{A.0.8}$$

Note that if $\boldsymbol{K}_1$ has full rank, we have due to Lemma 2.1.25

$$\boldsymbol{B}\boldsymbol{K}_1^{-1} = \boldsymbol{A}_2\boldsymbol{C}^{\frac{1}{2}}(\boldsymbol{A}_1\boldsymbol{C}^{\frac{1}{2}})^*\left(\boldsymbol{A}_1\boldsymbol{C}^{\frac{1}{2}}(\boldsymbol{A}_1\boldsymbol{C}^{\frac{1}{2}})^*\right)^{-1} = (\boldsymbol{A}_2\boldsymbol{C}^{\frac{1}{2}})(\boldsymbol{A}_1\boldsymbol{C}^{\frac{1}{2}})^\dagger$$

Therefore, Definition A.0.7 and Lemma A.0.2 yield

$$\left\|\boldsymbol{B}\boldsymbol{K}_1^{-1}\right\| \leq \sqrt{\|\boldsymbol{K}_2\|\left\|\boldsymbol{K}_1^\dagger\right\|} \leq \sqrt{\frac{3(\Delta N)^2 + \pi^2}{3(\Delta N)^2 - \pi^2}},$$

together with Lemma 2.1.22, we obtain

$$\left\|\begin{pmatrix}\boldsymbol{I}_{M/2} & \boldsymbol{0}_{M/2}\\ -\boldsymbol{B}\boldsymbol{K}_1^{-1} & \boldsymbol{I}_{M/2}\end{pmatrix}\right\|^2 \leq 1 + \left\|\boldsymbol{B}\boldsymbol{K}_1^{-1}\right\| + \left\|\boldsymbol{B}\boldsymbol{K}_1^{-1}\right\|^2 \leq \frac{6(\Delta N)^2}{3(\Delta N)^2 - \pi^2} + \sqrt{\frac{3(\Delta N)^2 + \pi^2}{3(\Delta N)^2 - \pi^2}}.$$

Now, we estimate $\left\|(\boldsymbol{K}_2 - \boldsymbol{B}\boldsymbol{K}_1^{-1}\boldsymbol{B}^*)^{-1}\right\|$, which is done by the following steps:

i) First, note that $\boldsymbol{I}_{M/2} - (\boldsymbol{A}_1\boldsymbol{C}^{\frac{1}{2}})^*(\boldsymbol{A}_1\boldsymbol{C}\boldsymbol{A}_1^*)^{-1}\boldsymbol{A}_1\boldsymbol{C}^{\frac{1}{2}} = \boldsymbol{I}_{M/2} - (\boldsymbol{A}_1\boldsymbol{C}^{\frac{1}{2}})^\dagger\boldsymbol{A}_1\boldsymbol{C}^{\frac{1}{2}}$ is an orthogonal projector and thus Theorem 3.1.7 implies

$$\left\|\boldsymbol{K}_2 - \boldsymbol{B}\boldsymbol{K}_1^{-1}\boldsymbol{B}^*\right\| \leq \left\|\boldsymbol{A}_2\boldsymbol{C}^{\frac{1}{2}}\right\|\left\|\boldsymbol{I}_{M/2} - \boldsymbol{A}_1^\dagger\boldsymbol{A}_1\right\|\left\|(\boldsymbol{A}_2\boldsymbol{C}^{\frac{1}{2}})^*\right\| \\ \leq \|\boldsymbol{K}_2\| \leq N\frac{2}{3}\left(1 + \frac{\pi^2}{12}\right) < 2N.$$

We apply Lemma 2.1.20 with $\eta = 2N$, use the identities $\boldsymbol{R}_1 = \boldsymbol{B} - \boldsymbol{K}_1$ and $\boldsymbol{R}_2 = \boldsymbol{B} - \boldsymbol{K}_2$, apply the triangular inequality, and the sub-multiplicativity of the matrix norm to get

$$\begin{aligned}\left\|(\boldsymbol{K}_2 - \boldsymbol{B}\boldsymbol{K}_1^{-1}\boldsymbol{B}^*)^{-1}\right\| &\leq \frac{1}{2N - \left\|2N\boldsymbol{I}_{M/2} - \boldsymbol{K}_2 + \boldsymbol{B}\boldsymbol{K}_1^{-1}\boldsymbol{B}^*\right\|} \\ &\leq \frac{1}{2N - \left\|2N\boldsymbol{I}_{M/2} + \boldsymbol{R}_1^* + \boldsymbol{R}_2\right\| - \|\boldsymbol{R}_1\|^2\left\|\boldsymbol{K}_1^{-1}\right\|}.\end{aligned} \tag{A.0.9}$$

ii) Lemma A.0.9 leads to

$$2N - \left\|2N\boldsymbol{I}_{M/2} + \boldsymbol{R}_1^* + \boldsymbol{R}_2\right\| \geq 2(n+1-F_n(q)) - c^2(n+1)^3q^2 \cdot \left(\frac{\pi^4}{3((n+1)\Delta)^2} + \frac{4.84\pi}{((n+1)\Delta)^3} + \frac{8\pi^4+\pi^6}{360((n+1)\Delta)^4}\right).$$

iii) We apply Theorem A.0.2 and Lemma A.0.8 to get

$$\|\boldsymbol{R}_1\|^2 \left\|\boldsymbol{K}_1^{-1}\right\| \leq \frac{1}{(n+1)} \frac{3(N\Delta)^2}{3(N\Delta)^2-\pi^2} \cdot \left[n+1-F_n(cq) + c(n+1)^2 q\left(\frac{\pi^3}{6((n+1)\Delta)^2} + \frac{1.21}{((n+1)\Delta)^3}\right)\right]^2.$$

iv) We use the estimates for the Fejér kernel $n+1-F_n(q) \geq (n+1)^3q^2$ in ii) and $n+1-F_n(cq) \leq \frac{c^2\pi^2}{3}(n+1)^3q^2$ in iii), see Lemma A.0.1, and $\frac{n+1}{N} \leq \frac{2}{3}$. Inserting this in (A.0.9) finally yields

$$\begin{aligned}
&\left\|(\boldsymbol{K}_2 - \boldsymbol{B}\boldsymbol{K}_1^{-1}\boldsymbol{B}^*)^{-1}\right\| \\
&\leq \frac{1}{(n+1)^3q^2}\left[2 - \frac{c^2\pi^4}{3((n+1)\Delta)^2} - \frac{4.84c^2\pi}{((n+1)\Delta)^3} - \frac{8c^2\pi^4+c^2\pi^6}{360((n+1)\Delta)^4}\right. \\
&\quad \left. - \frac{3(N\Delta)^2}{3(N\Delta)^2-\pi^2} \cdot \left(\frac{c^2\pi^2}{3}(n+1)q + \frac{c\pi^3}{6((n+1)\Delta)^2} + \frac{1.21c}{((n+1)\Delta)^3}\right)^2\right]^{-1} \\
&\leq \frac{8}{N(qN)^2}\left[2 - \frac{4c^2\pi^4}{3(\Delta N)^2} - \frac{39c^2\pi}{(\Delta N)^3} - \frac{16c^2\pi^4+2c^2\pi^6}{45(\Delta N)^4}\right. \\
&\quad \left. - \frac{3(\Delta N)^2}{3(\Delta N)^2-\pi^2}\left(\frac{2c^2\pi^2}{9}qN + \frac{2c\pi^3}{3(\Delta N)^2} + \frac{9.68c}{(\Delta N)^3}\right)^2\right]^{-1}.
\end{aligned}$$

This upper bound also bounds the maximum in (A.0.8) since for all $qN \leq 1/2$ and $\Delta N \geq 2$ together with Lemma A.0.2

$$\left\|\boldsymbol{K}_1^{-1}\right\| \leq \frac{2}{N} \cdot \frac{12}{12-\pi^2} \leq \frac{8}{N(qN)^2\frac{8}{6}(12-\pi^2)} \leq \frac{8}{N(qN)^2}[2-\cdots]^{-1}.$$

□

Theorem A.0.11 (Upper bound)**.**
Under the conditions of (3.3.9) *with* $qN \leq q_{\max}N = \frac{1}{4c^2}$ *and* $\Delta N \geq \Delta_{\min}N = 11c^2$*, we have*

$$\operatorname{cond}(\boldsymbol{A}) \leq \frac{14}{qN}.$$

Proof. In Lemma A.0.10 the constant $C(qN, \Delta N, c)$ is monotone increasing in qN and monotone decreasing in ΔN. Hence, after plugging in the bounds for qN and ΔN in our assumptions, it is easy to see that the constant $C(\frac{1}{4c^2}, 10c^2, c)$ is monotone decreasing in c. Therefore, we get $C(qN, \Delta N, c) \leq C(1/4, 10, 1) \leq 86$, so that $\left\|\widetilde{\boldsymbol{K}}^{-1}\right\| \leq 86N^{-1}(qN)^{-2}$. Together with Lemma 2.1.17 and the bound $\|\boldsymbol{K}\| \leq 22N/10 = 2.2N$ from Lemma 3.3.9 we obtain the result. □

If the uniformity parameter $c = 1$ in (3.3.9) then we get the following results.

Lemma A.0.12.
Under the conditions of (3.3.9) *with* $c = 1$*, such that*

$$\widetilde{C}(\Delta N) := \frac{1}{8}\Bigg(2 - \frac{4\pi^4}{3(\Delta N)^2} - \frac{39\pi}{(\Delta N)^3} - \frac{16\pi^4 + 2\pi^6}{45(\Delta N)^4} - \frac{3(\Delta N)^2}{3(\Delta N)^2 - \pi^2} \\ - \frac{3(\Delta N)^2}{3(\Delta N)^2 - \pi^2}\left(\frac{8\pi^3}{9(\Delta N)^2} + \frac{13}{(\Delta N)^3}\right) - \frac{3(\Delta N)^2}{3(\Delta N)^2 - \pi^2}\left(\frac{2\pi^3}{3(\Delta N)^2} + \frac{9.68}{(\Delta N)^3}\right)^2\Bigg)$$

is positive, we have

$$\left\|K^{-1}\right\| \leq \frac{C(\Delta N)}{N(qN)^2},$$

where $C(\Delta N) := \left(\frac{6(\Delta N)^2}{3(\Delta N)^2 - \pi^2} + \sqrt{\frac{3(\Delta N)^2 + \pi^2}{3(\Delta N)^2 - \pi^2}}\right)/\widetilde{C}(\Delta N)$.

Proof. The proof is analogous to that of Lemma A.0.10, the only difference is in step iv). Setting $c = 1$ in ii) and iii), expanding the squared bracket in iii) and inserting this into (A.0.9) leads to

$$\begin{aligned}\left\|(\boldsymbol{K}_2 - \boldsymbol{B}\boldsymbol{K}_1^{-1}\boldsymbol{B}^*)^{-1}\right\| \leq \Bigg[&2(n+1-F_n(q)) \\ &-(n+1)^3q^2 \cdot \left(\frac{\pi^4}{3((n+1)\Delta)^2} + \frac{4.84\pi}{((n+1)\Delta)^3} + \frac{8\pi^4 + \pi^6}{360((n+1)\Delta)^4}\right) \\ &-\frac{1}{n+1} \cdot \frac{3(N\Delta)^2}{3(N\Delta)^2 - \pi^2} \cdot \Bigg[(n+1-F_n(q))^2 \\ &\quad + (n+1-F_n(q))2(n+1)^2q\left(\frac{\pi^3}{6((n+1)\Delta)^2} + \frac{1.21}{((n+1)\Delta)^3}\right) \\ &\quad + (n+1)^4q^2\left(\frac{\pi^3}{6((n+1)\Delta)^2} + \frac{1.21}{((n+1)\Delta)^3}\right)^2\Bigg]\Bigg]^{-1}.\end{aligned}$$

In three summands, we can factor out $n + 1 - F_n(q)$ and use the estimate $n + 1 - F_n(q) \geq (n+1)^3q^2$. Additionally, in the third summand we use the rough bound $n+1-F_n(q) \leq n+1$ and in the fourth $(n + 1)q \leq \frac{2}{3}$. We obtain

$$\begin{aligned}\left\|(\boldsymbol{K}_2 - \boldsymbol{B}\boldsymbol{K}_1^{-1}\boldsymbol{B}^*)^{-1}\right\| &\leq \frac{1}{(n+1)^3q^2}\Bigg[2 - \frac{\pi^4}{3((n+1)\Delta)^2} - \frac{4.84\pi}{((n+1)\Delta)^3} - \frac{8\pi^4 + \pi^6}{360((n+1)\Delta)^4} \\ &\quad - \frac{3(\Delta N)^2}{3(\Delta N)^2 - \pi^2} - \frac{3(\Delta N)^2}{3(\Delta N)^2 - \pi^2}\left(\frac{2\pi^3}{9((n+1)\Delta)^2} + \frac{1.62}{((n+1)\Delta)^3}\right) \\ &\quad - \frac{3(\Delta N)^2}{3(\Delta N)^2 - \pi^2}\left(\frac{\pi^3}{6((n+1)\Delta)^2} + \frac{1.21}{((n+1)\Delta)^3}\right)^2\Bigg]^{-1} \\ &\leq \frac{8}{N(\Delta N)^2}\Bigg[2 - \frac{4\pi^4}{3(\Delta N)^2} - \frac{39\pi}{(\Delta N)^3} - \frac{16\pi^4 + 2\pi^6}{45(\Delta N)^4} - \frac{3(\Delta N)^2}{3(\Delta N)^2 - \pi^2} \\ &\quad - \frac{3(\Delta N)^2}{3(\Delta N)^2 - \pi^2}\left(\frac{8\pi^3}{9(\Delta N)^2} + \frac{13}{(\Delta N)^3}\right)\end{aligned}$$

$$- \frac{3(\Delta N)^2}{3(\Delta N)^2 - \pi^2} \left(\frac{2\pi^3}{3(\Delta N)^2} + \frac{9.68}{(\Delta N)^3} \right)^2 \Bigg]^{-1}$$

The same argument as in (A.0.6) shows that this also bounds the maximum in A.0.8 and we get the result. □

Theorem A.0.13 (Upper bound)**.**
Under the conditions of (3.3.9) *with* $c = 1$, $\Delta N \geq \Delta_{\min} N = 25$, *we have*

$$\operatorname{cond}(\boldsymbol{A}) < \frac{8.3}{qN}.$$

Proof. Direct inspection gives monotonicity of $C(\Delta)$ with respect to Δ and also the estimate $C(25/N) \leq 33$. Hence $\left\| \hat{\boldsymbol{K}}^{-1} \right\| \leq 33 N^{-1} (qN)^{-2}$ and together with Lemma 2.1.17 and the bound $\|\boldsymbol{K}\| \leq 52N/25$ from Lemma 3.3.9 we obtain the result. □

Appendix B

QR method for pair clusters

Here we present the application of the QR-decomposition technique for estimating the extremal singular values of Vandermonde matrices with node set that consists of pair clusters. The results are referred to in Remark 3.4.26.

Let $N = 2n + 1$ for some $n \in \mathbb{N}$, $M = 4$ and $\boldsymbol{A} \in \mathbb{C}^{M\times N}$ be the Vandermonde matrix corresponding to the nodes

$$t_1 = 0 < t_2 = t_1 + q_1 \ll t_3 = t_2 + \Delta < t_4 = t_3 + q_2 \ll 1 \tag{B.0.1}$$

in $\mathbb{T}$ with $0 < q_1 \leq q_2 \leq 1/N$ and $\Delta > 1/N$. This is a node set of two pair clusters, see (3.3.9). According to the cluster structure denote $\boldsymbol{A}_1$ and $\boldsymbol{A}_2$ the submatrices of $\operatorname{diag}(\mathrm{e}^{-2\pi i n t_1}, \ldots, \mathrm{e}^{-2\pi i n t_4})\boldsymbol{A}$ (the shifted Vandermonde matrix) by taking the first and the second two rows, respectively.

Lemma B.0.1.
There exist upper triangular matrices $\boldsymbol{R}_1, \boldsymbol{R}_2 \in \mathbb{C}^{2\times 2}$ such that the QR-decompositions of the submatrices are given by $\boldsymbol{A}_1^* = \boldsymbol{Q}_1\boldsymbol{R}_1$ and $\boldsymbol{A}_2^* = \boldsymbol{Q}_2\boldsymbol{R}_2$, with

$$\begin{aligned}
\boldsymbol{Q}_1 &= \left(1/\sqrt{N}, \tfrac{\mathrm{e}^{-2\pi i k q_1} - D_n(q_1)/N}{\sqrt{N - D_n(q_1)^2/N}}\right)_{|k|\leq n}, \\
\boldsymbol{Q}_2 &= \operatorname{diag}(\mathrm{e}^{-2\pi i k q_2})_{|k|\leq n} \cdot \left(1/\sqrt{N}, \tfrac{\mathrm{e}^{-2\pi i k q_2} - D_n(q_2)/N}{\sqrt{N - D_n(q_2)^2/N}}\right)_{|k|\leq n}.
\end{aligned}$$

Proof. We orthnormalize the columns of $\boldsymbol{A}_1^*$ by a Gram-Schmidt step. The first column becomes

$$(\boldsymbol{Q}_1)_{:,1} = \frac{1}{\sqrt{N}}(\boldsymbol{A}_1^*)_{:,1}$$

and the second $\boldsymbol{v}/\left\|\boldsymbol{v}\right\|$, where

$$\boldsymbol{v} := (\boldsymbol{A}_1^*)_{:,2} - \langle (\boldsymbol{A}_1^*)_{:,2}, (\boldsymbol{Q}_1)_{:,1}\rangle (\boldsymbol{Q}_1)_{:,1} = (\boldsymbol{A}_1^*)_{:,2} - \frac{1}{N}\left(D_n(q_1)\right)_{|k|\leq 1}.$$

Now we have

$$\left\|\boldsymbol{v}\right\|^2 = \sum_{k=-n}^{n} \left(\mathrm{e}^{2\pi i k q_1} - \frac{D_n(q_1)}{N}\right)\left(\mathrm{e}^{-2\pi i k q_1} - \frac{D_n(q_1)}{N}\right)$$

$$
\begin{aligned}
&= \sum_{k=-n}^{n} \left(1 - \frac{D_n(q_1)}{N} \left(\mathrm{e}^{2\pi i k q_1} + \mathrm{e}^{-2\pi i k q_1} \right) + \frac{D_n(q_1)^2}{N^2} \right) \\
&= N - \frac{2D_n(q_1)^2}{N} + \frac{D_n(q_1)^2}{N} \\
&= \frac{N^2 - D_n(q_1)^2}{N}
\end{aligned}
$$

and therefore

$$
\boldsymbol{A}_1^* = \boldsymbol{Q}_1 \boldsymbol{R}_1 = \left(1/\sqrt{N}, \quad \frac{\mathrm{e}^{-2\pi i k q_1} - D_n(q_1)/N}{\sqrt{N - D_n(q_1)^2/N}} \right)_{|k| \le n} \cdot \begin{pmatrix} \sqrt{N} & \frac{D_n(q_1)}{\sqrt{N}} \\ 0 & \|\boldsymbol{v}\| \end{pmatrix}.
$$

For $\boldsymbol{A}_2^*$, notice that $\boldsymbol{A}_2^* = \mathrm{diag}(\mathrm{e}^{-2\pi i k t_3})_{|k| \le n} \cdot \left(1, \mathrm{e}^{-2\pi i k q_2}\right)_{|k| \le n}$ and proceed with the right hand matrix analogously to $\boldsymbol{A}_1^*$. □

Lemma B.0.2.
For $0 \le q_1 \le q_2 \le \frac{1}{N} \le 1$ and $\ell \ge 2$ we have,

$$
\sum_{r=1}^{\ell-1} \binom{\ell}{r} q_2^r (-q_1)^{\ell - r} \le q_1 q_2 \frac{\ell}{N^{\ell-2}}.
$$

Proof. Let $q_1 = b q_2$ for some $0 \le b \le 1$. Then it holds

$$
\begin{aligned}
\sum_{r=1}^{\ell-1} \binom{\ell}{r} q_2^r (-q_1)^{\ell - r} &\le (q_2 - q_1)^\ell - q_2^\ell - (-q_1)^\ell \\
&= (q_2 - b q_2)^\ell - q_2^\ell - (-b q_2)^\ell \\
&= q_2^\ell \left((1-b)^\ell - 1 - (-b)^\ell \right) \\
&= q_1 q_2 \frac{1}{N^{\ell-2}} \frac{1}{b} \left((1-b)^\ell - 1 - (-b)^\ell \right).
\end{aligned}
$$

We still have to show that $\frac{1}{b}\left((1-b)^\ell - 1 - (-b)^\ell \right) \le \ell$. This is equivalent to

$$
(1-b)^\ell \le 1 + \ell b + (-b)^\ell.
$$

For ℓ even, it is a weaker form of the Bernoulli inequality. Then we use this for the case ℓ odd and obtain

$$
\begin{aligned}
(1-b)^\ell &= (1-b)(1-b)^{\ell-1} \\
&\le (1-b) \left(1 + (\ell - 1) b + (-b)^{\ell-1} \right) \\
&= 1 + \ell b + (-b)^\ell - 2b + (-b)^{\ell-1} - (\ell - 1) b^2 \\
&= 1 + \ell b + (-b)^\ell + b \left(-2 + b^{\ell-2} - (\ell - 1) b \right) \\
&\le 1 + \ell b + (-b)^\ell.
\end{aligned}
$$

□

Lemma B.0.3.
Under the assumptions of Lemma B.0.1, it holds

$$\|\boldsymbol{Q}_1^*\boldsymbol{Q}_2\|_F \leq \frac{116}{\Delta N}.$$

Proof. Throughout the proof we use the bounds

$$N - N^2 t \leq D_n(t) \leq N - N^3 t^2, \quad \text{for } 0 \leq t \leq \frac{1}{N}$$

and

$$|D_n(t)| \leq \frac{1}{2t}, \qquad \left|D_n^{(\ell)}(t)\right| \leq \frac{(\pi N)^\ell}{t}, \quad \text{for } 0 < t \leq \frac{1}{2}$$

from Lemmata 2.4.7 and 2.4.9. We bound the absolute values of the entries of $\boldsymbol{Q}_1^*\boldsymbol{Q}_2$.

$$|(\boldsymbol{Q}_1^*\boldsymbol{Q}_2)_{1,1}| = \frac{|D_n(\Delta)|}{N} \leq \frac{1}{2\Delta N}.$$

For some $\xi_1 \in (t_3, t_3 + q_2)$, i.e. $\xi_1 \geq \Delta$, we have

$$\begin{aligned}
|(\boldsymbol{Q}_1^*\boldsymbol{Q}_2)_{1,2}| &= \left| \frac{1}{\sqrt{N^2 - D_n(q_2)^2}} \left(D_n(t_3 + q_2) - \frac{D_n(q_2)}{N} D_n(t_3) \right) \right| \\
&\leq \frac{|D_n(t_3 + q_2) - D_n(t_3)| + \frac{D_n(t_3)}{N}\left[N - D_n(q_2)\right]}{\sqrt{|N + D_n(q_2)|\,|N - D_n(q_2)|}} \\
&\leq \frac{1}{q_2 N^2} \left(\left|D_n'(\xi_1)\right| q_2 + \frac{|D_n(t_3)|\, q_2 N^2}{N} \right) \\
&\leq \frac{\pi + 1/2}{\Delta N} \leq \frac{3.7}{\Delta N}
\end{aligned}$$

Analogously, we get the same bound

$$|(\boldsymbol{Q}_1^*\boldsymbol{Q}_2)_{2,1}| \leq \frac{3.7}{\Delta N}.$$

For the last entry, due to the Taylor expansion of $D_n(t)$ at t_3, the bound $q_1, q_2 \leq 1/N$ and the bound from Lemma B.0.2, we have

$$\begin{aligned}
&|(\boldsymbol{Q}_1^*\boldsymbol{Q}_2)_{2,2}| = \left| \frac{N}{\sqrt{N^2 - D_n(q_1)^2}\sqrt{N^2 - D_n(q_2)^2}} \right| \\
&\left| \left(D_n(t_3 - q_1 + q_2) - \frac{D_n(q_2)}{N} D_n(t_3 - q_1) - \frac{D_n(q_1)}{N} D_n(t_3 + q_2) + \frac{D_n(q_1) D_n(q_2)}{N^2} D_n(t_3) \right) \right| \\
&\leq \frac{1}{N^3 q_1 q_2} \Bigg[\left| \frac{D_n(t_3)}{N^2} (N - D_n(q_2))(N - D_n(q_1)) \right| \\
&\quad + \left| D_n'(t_3) \left((q_2 - q_1) - \frac{D_n(q_2)}{N}(-q_1) - \frac{D_n(q_1)}{N} q_2 \right) \right| \\
&\quad + \left| \sum_{\ell=2}^{\infty} \frac{D_n^{(\ell)}(t_3)}{\ell!} \left((q_2 - q_1)^\ell - \frac{D_n(q_2)}{N}(-q_1)^\ell - \frac{D_n(q_1)}{N} q_2^\ell \right) \right| \Bigg]
\end{aligned}$$

$$
\begin{aligned}
&\leq \frac{1}{2\Delta N} + \frac{\pi}{\Delta N^2 q_1 q_2} \left| \frac{q_2}{N} \left(N - D_n\left(q_1\right)\right) - \frac{q_1}{N} \left(N - D_n\left(q_2\right)\right) \right| \\
&\quad + \left| \sum_{\ell=2}^{\infty} \frac{D_n^{(\ell)}\left(t_3\right)}{\ell! N^3 q_1 q_2} \left(\frac{q_2^\ell}{N} \left(N - D_n\left(q_1\right)\right) + \frac{(-q_1)^\ell}{N} \left(N - D_n\left(q_2\right)\right) + \sum_{r=1}^{\ell-1} \binom{\ell}{r} q_2^r (-q_1)^{\ell-r} \right) \right| \\
&\leq \frac{1}{\Delta N} + \frac{2\pi}{\Delta N} + \sum_{\ell=2}^{\infty} \frac{\left|D_n^{(\ell)}\left(t_3\right)\right|}{\ell! N^3 q_1 q_2} \frac{2 q_1 q_2}{N^{\ell-2}} + \left| \sum_{\ell=2}^{\infty} \frac{D_n^{(\ell)}\left(t_3\right)}{\ell! N^3 q_1 q_2} \sum_{r=1}^{\ell-1} \binom{\ell}{r} q_2^r (-q_1)^{\ell-r} \right| \\
&\leq \frac{1}{2\Delta N} + \frac{2\pi}{\Delta N} + \frac{2}{\Delta N} \sum_{\ell=2}^{\infty} \frac{\pi^\ell}{\ell!} + \frac{1}{\Delta N} \sum_{\ell=2}^{\infty} \frac{\pi^\ell N^{\ell-2}}{\ell! q_1 q_2} \left| \sum_{r=1}^{\ell-1} \binom{\ell}{r} q_2^r (-q_1)^{\ell-r} \right| \\
&\leq \frac{1}{2\Delta N} + \frac{2\pi}{\Delta N} + \frac{2}{\Delta N} (\mathrm{e}^\pi - \pi - 1) + \frac{1}{\Delta N} \sum_{\ell=2}^{\infty} \frac{\pi^\ell}{(\ell-1)!} \\
&= \frac{1}{2\Delta N} + \frac{2\pi}{\Delta N} + \frac{2}{\Delta N} (\mathrm{e}^\pi - \pi - 1) + \frac{1}{\Delta N} (\mathrm{e}^\pi - 1)\pi \\
&= \frac{1}{\Delta N} \left(\frac{1}{2} + 2\pi + 2(\mathrm{e}^\pi - \pi - 1) + \pi(\mathrm{e}^\pi - 1) \right) \\
&< \frac{115}{\Delta N}.
\end{aligned}
$$

Using the 4 upper bounds in the Frobenius norm of $\boldsymbol{Q}_1^* \boldsymbol{Q}_2$ yields the result. □

Theorem B.0.4.
Let $N \in \mathbb{N}_+$ be odd and $\boldsymbol{A}$ be Vandermonde matrix as in (3.1.6) *with nodes consisting of pair clusters as in* (3.3.9)*. Then we have*

$$
qN^{3/2} \sqrt{1 - \frac{232(\log(\lfloor M/4 \rfloor) + 1)}{\Delta N}} \leq \sigma_{\min}(\boldsymbol{A}) \leq \sigma_{\max}(\boldsymbol{A}) \leq \sqrt{2N} \sqrt{1 + \frac{232(\log(\lfloor M/4 \rfloor) + 1)}{\Delta N}}.
$$

Proof. Without loss of generality we deal with the shifted Vandermonde $\check{\boldsymbol{A}}$ and follow the proof of [12, Lem. 5.1]. Therefore, we partition $\boldsymbol{A}$ according to the clusters and their corresponding rows into the matrices $\boldsymbol{A}_1, \ldots, \boldsymbol{A}_{M/2}$. For each of these matrices we apply a QR decomposition as in Lemma B.0.1 such that $\boldsymbol{A}_j^* = \boldsymbol{Q}_j \boldsymbol{R}_j$. Now define the matrix $\boldsymbol{Q} := \left(\boldsymbol{Q}_1, \cdots, \boldsymbol{Q}_{M/2}\right)$ and the block diagonal matrix $\boldsymbol{R} := \operatorname{diag}(\boldsymbol{R}_1, \ldots, \boldsymbol{R}_{M/2})$. Note that $\boldsymbol{A}^* = \boldsymbol{Q}\boldsymbol{R}$ and $\operatorname{range}(\boldsymbol{Q}) = \operatorname{range}(\boldsymbol{A}^*)$. Furthermore, since $\sigma_r(\boldsymbol{A}_j^*) = \sigma_r(\boldsymbol{R}_j)$ for all r and j, $\boldsymbol{R}$ has the collection of all singular values of the $\boldsymbol{A}_j$ as own singular values. Moreover, as a consequence of Lemma 2.1.16, we have

$$
\sigma_{\min}(\boldsymbol{Q})\sigma_j(\boldsymbol{R}) \leq \sigma_j(\boldsymbol{A}) \leq \sigma_{\max}(\boldsymbol{Q})\sigma_j(\boldsymbol{R}), \quad 1 \leq j \leq M, \tag{B.0.2}
$$

which links the singular values of the cluster-corresponding Vandermonde matrices $\boldsymbol{A}_j$ to the singular values of the whole Vandermonde matrix $\boldsymbol{A}$. Now, if q_j denotes the distance between the nodes of the cluster corresponding to $\boldsymbol{A}_j$, then their singular values are explicitly given by

$$
\sigma_{\min}(\boldsymbol{A}_j)^2 = \lambda_{\min}(\boldsymbol{A}_j \boldsymbol{A}_j^*) = \lambda_{\min} \begin{pmatrix} N & D_n\left(q_j\right) \\ D_n\left(q_j\right) & N \end{pmatrix} = N - D_n\left(q_j\right) \geq q_j^2 N^3 \tag{B.0.3}
$$

and

$$\sigma_{\max}(\boldsymbol{A}_j)^2 = \lambda_{\max}(\boldsymbol{A}_j\boldsymbol{A}_j^*) = \lambda_{\max}\begin{pmatrix} N & D_n(q_j) \\ D_n(q_j) & N \end{pmatrix} = N + D_n(q_j) \leq 2N. \tag{B.0.4}$$

Especially, we obtain $\sqrt{N}qN \leq \sigma_{\min}(\boldsymbol{R}) < \sigma_{\max}(\boldsymbol{R}) \leq \sqrt{2N}$. Finally, bounds for $\sigma_{\min}(\boldsymbol{Q})$ and $\sigma_{\max}(\boldsymbol{Q})$ are established by looking at

$$\boldsymbol{Q}^*\boldsymbol{Q} = \begin{pmatrix} \boldsymbol{Q}_1^*\boldsymbol{Q}_1 & \boldsymbol{Q}_1^*\boldsymbol{Q}_2 & \cdots & \boldsymbol{Q}_1^*\boldsymbol{Q}_{M/2} \\ \boldsymbol{Q}_2^*\boldsymbol{Q}_1 & \boldsymbol{Q}_2^*\boldsymbol{Q}_2 & \cdots & \boldsymbol{Q}_2^*\boldsymbol{Q}_1 \\ \vdots & \vdots & \vdots & \vdots \\ \boldsymbol{Q}_{M/2}^*\boldsymbol{Q}_1 & \boldsymbol{Q}_{M/2}^*\boldsymbol{Q}_2 & \cdots & \boldsymbol{Q}_{M/2}^*\boldsymbol{Q}_{M/2} \end{pmatrix}.$$

Since $\boldsymbol{Q}_j$ are orthogonal, we write $\boldsymbol{Q}^*\boldsymbol{Q} = \boldsymbol{I}_M + \boldsymbol{E}$, where $\boldsymbol{I}_M$ denotes the identity matrix in $\mathbb{C}^{M\times M}$ and $\boldsymbol{E}$ is a hermitian matrix consisting of the off block-diagonal entries $\langle \boldsymbol{u}, \boldsymbol{v}\rangle$, with $\boldsymbol{u}$ being a column of $\boldsymbol{Q}_j$, $\boldsymbol{v}$ being a column of $\boldsymbol{Q}_k$, for $j \neq k$. These entries can be bound according to Lemma B.0.3. Using the minimal cluster separation, the packing argument, that at most two clusters are Δ close to one cluster and then at most two clusters 2Δ close to it and so on, leads to an entry-wise upper bounding, symmetric matrix. Therefore, denoting the minimal distance of nodes in clusters corresponding to $\boldsymbol{A}_j$ and $\boldsymbol{A}_k$ by Δ_{jk}, Lemma 2.1.23 and Lemma 2.1.22 provide

$$\begin{aligned} \|\boldsymbol{E}\| &\leq \max_{1\leq j\leq M/2} \sum_{k=1,k\neq j}^{M/2} \left\|\boldsymbol{Q}_j^*\boldsymbol{Q}_k\right\|_F \leq \max_{1\leq j\leq M/2} \sum_{k=1,k\neq j}^{M/2} \frac{116}{\Delta_{jk}N} \\ &\leq 2\sum_{k=1,k\neq j}^{\lfloor M/4\rfloor} \frac{116}{k\Delta N} \leq \frac{232(\log(\lfloor \frac{M}{4}\rfloor)+1))}{\Delta N}. \end{aligned}$$

We have $1 - \|\boldsymbol{E}\| \leq \lambda_{\min}(\boldsymbol{Q}^*\boldsymbol{Q}) \leq 1 + \|\boldsymbol{E}\|$ by Weyl's perturbation inequality, so that altogether, we get

$$\begin{aligned} \sigma_{\min}(\boldsymbol{Q}) &= \sqrt{\lambda_{\min}(\boldsymbol{Q}^*\boldsymbol{Q})} \geq \sqrt{1 - \frac{232(\log(\lfloor \frac{M}{4}\rfloor)+1)}{\Delta N}} \\ \sigma_{\max}(\boldsymbol{Q}) &= \sqrt{\lambda_{\max}(\boldsymbol{Q}^*\boldsymbol{Q})} \leq \sqrt{1 + \frac{232(\log(\lfloor \frac{M}{4}\rfloor)+1)}{\Delta N}} \end{aligned}$$

and with (B.0.2), (B.0.3) and (B.0.4) the result. □

Theorem B.0.5.
Under the assumptions of Theorem B.0.4 by means of [12] we get the bounds

$$\begin{aligned} \sigma_{\min}(\boldsymbol{A}) &\geq \sqrt{N}qN\sqrt{1 - \left(\frac{141M}{\Delta(N-1)} + 1079MqN\right)}, \\ \sigma_{\max}(\boldsymbol{A}) &\geq \sqrt{2N}\sqrt{1 + \left(\frac{141M}{\Delta(N-1)} + 1079MqN\right)}. \end{aligned}$$

Proof. The main steps of the technique developed in [12] are used in the proof of Theorem B.0.4. The difference is that in [12] the entries in $\boldsymbol{Q}_j^*\boldsymbol{Q}_k$, for $j \neq k$, are upper bounded by means of the minimal angle between subspaces spanned by the respective column vectors of $\boldsymbol{Q}_j$ and $\boldsymbol{Q}_k$. If the smallest angle between pairs of those subspaces is $\frac{\pi}{2} - \alpha$, then [12, Lem. 5.1, p. 60] provides

$$\sqrt{1 - M\alpha} \leq \sigma_{\min}(\boldsymbol{Q}) \leq \sigma_{\max}(\boldsymbol{Q}) \leq \sqrt{1 + M\alpha}.$$

In [12, Prop. 5.1, p. 62] this parameter α gets upper bounded (under several technical conditions) by

$$\alpha \leq \frac{C_{25}}{2\pi\Delta(N-1)} + C_{26}2\pi q(N-1) \quad \left(\leq \frac{1}{M}\right).$$

Since we are only looking at pair clusters the constants C_{28} and C_{29}, which are only dependent on the number of nodes in the clusters, become explicit. We have the following forward list of constants development in [12]:

- $C_{25} = C_4$ and $C_{26} = C_3$ [p. 63],
- $C_4 = \frac{\pi}{2}C_{20}$ and $C_3 = \frac{\pi}{2}C_{21}$ [proof of Thm. 2.1, p. 54],
- $C_{20} = 8\Xi^{-2}C_{19}$ and $C_{21} = 4C_{22}\Xi^{-2}$ [p. 55],
- $C_{19} = 3\pi$ [p. 53],
- $C_{22} = 2C_{18}(2 + C_{18})$ [p. 55],
- $C_{18} = 2\sqrt{2}$ [p. 49].

The quantity Ξ is basically the square root of the smallest eigenvalue of a normalized Hilbert matrix, which is controlled asymptotically and serves as lower bound for $\sigma_{\min}(\boldsymbol{U}(\xi, N, 2))$, where $\boldsymbol{U}$ is the normalized limit bases matrix corresponding to a cluster containing ξ. In our case we have

$$\boldsymbol{U}(\xi, N, 2) = \begin{pmatrix} 1 & 0 \\ \mathrm{e}^{2\pi i\xi} & i\mathrm{e}^{2\pi i\xi} \\ \mathrm{e}^{2\pi i2\xi} & 2i\mathrm{e}^{2\pi i2\xi} \\ \vdots & \vdots \\ \mathrm{e}^{2\pi i(N-1)\xi} & (N-1)i\mathrm{e}^{2\pi i(N-1)\xi} \end{pmatrix} \cdot \begin{pmatrix} \frac{1}{\sqrt{N}} & 0 \\ 0 & \frac{1}{\sqrt{\frac{1}{6}(N-1)N(2N-1)}} \end{pmatrix}.$$

Hence, we can bound $\sigma_{\min}(\boldsymbol{U}(\xi, N, 2))$ directly instead of using Ξ. We obtain

$$\sigma_{\min}(\boldsymbol{U})^2 = \lambda_{\min}(\boldsymbol{U}^*\boldsymbol{U}) = \lambda_{\min}\begin{pmatrix} 1 & \sqrt{\frac{3N-3}{4N-2}} \\ \sqrt{\frac{3N-3}{4N-2}} & 1 \end{pmatrix} = 1 - \sqrt{\frac{3N-3}{4N-2}} \geq 1 - \frac{\sqrt{3}}{2}.$$

Going through the above list from bottom to top, we finally get

$$C_{28} = \frac{12\pi^2}{1 - \frac{\sqrt{3}}{2}}, \quad C_{29} = \frac{8\sqrt{2}(2 + 2\sqrt{2})\pi}{1 - \frac{\sqrt{3}}{2}}$$

and

$$\alpha \leq \frac{6\pi}{(1 - \frac{\sqrt{3}}{2})\Delta(N-1)} + \frac{16\pi^2\sqrt{2}(2 + 2\sqrt{2})qN}{1 - \frac{\sqrt{3}}{2}} \leq \frac{141}{\Delta(N-1)} + 1079qN.$$

□

Appendix C

Stability of ESPRIT – Comparable results

We use the same notation as in Chapter 4. In [5] a different approach is chosen to analyze the stability of the ESPRIT algorithm. Here we present the steps that are done to adjust the results in there for the comparison in Remark 4.3.6. The main differences in [5] are, firstly, that [5] does not make use of the principal vector basis, instead a step of the proof of Theorem 2 in [116, eq. (4)] is adapted to show with Wedin's theorem that there exists a unitary matrix $\boldsymbol{Q} \in \mathbb{C}^{M\times M}$ such that

$$\left\|\widetilde{\boldsymbol{U}}\boldsymbol{Q} - \boldsymbol{U}\right\|_{\mathrm{F}} \leq \frac{\sqrt{2}\,\|\boldsymbol{E}\|_{\mathrm{F}}}{\sigma_M(\boldsymbol{H}_L) - \|\boldsymbol{E}\|}. \tag{C.0.1}$$

Secondly, the perturbation results for the pseudo inverse are not used. Instead, a perturbation result for least squares problems from [58, Fact 5.14] modified for application to matrices rather than vectors is used, [5, p. 199, Prop. 12]. The details can be found in [5]. Here we just apply them and add slightly stronger assumption as we did in Theorem 4.3.5 to obtain simplified and comparable results.

Theorem 2.2.3 in connection with the above matrix $\boldsymbol{Q}$ used as similar transformation for $\widetilde{\boldsymbol{X}}$ and Lemma 4.3.1 yields

$$\mathrm{md}_{\mathbb{T}}(\Omega, \widetilde{\Omega}) \leq \frac{1}{2}(2M-1)\frac{\|\boldsymbol{A}_L\|}{\sigma_{\min}(\boldsymbol{A}_L)}\left\|\boldsymbol{Q}^*\widetilde{\boldsymbol{X}}\boldsymbol{Q} - \boldsymbol{X}\right\|. \tag{C.0.2}$$

Since the matrix $\widetilde{\boldsymbol{X}} = \widetilde{\boldsymbol{U}}_{\uparrow}^{\dagger}\widetilde{\boldsymbol{U}}_{\downarrow}$, we have

$$\boldsymbol{Q}^*\widetilde{\boldsymbol{X}}\boldsymbol{Q} = \left(\widetilde{\boldsymbol{U}}_{\uparrow}\boldsymbol{Q}\right)^{\dagger}\left(\widetilde{\boldsymbol{U}}_{\downarrow}\boldsymbol{Q}\right) = \left(\boldsymbol{U}_{\uparrow} + (\widetilde{\boldsymbol{U}}_{\uparrow}\boldsymbol{Q} - \boldsymbol{U}_{\uparrow})\right)^{\dagger}\left(\boldsymbol{U}_{\downarrow} + (\widetilde{\boldsymbol{U}}_{\downarrow}\boldsymbol{Q} - \boldsymbol{U}_{\downarrow})\right).$$

Hence, this can be regarded as a solution to the least Frobenius norm problem

$$\min_{\boldsymbol{M}\in\mathbb{C}^{M\times M}} \|\boldsymbol{U}_{\uparrow}\boldsymbol{M} - \boldsymbol{U}_{\downarrow}\|_{\mathrm{F}}$$

with additional perturbations (Matrix analogue to the least squares problem). Proposition 12 from [5, p. 199] then provides a bound on the distance between the solution matrix $\boldsymbol{Q}^*\widetilde{\boldsymbol{X}}\boldsymbol{Q}$ of the perturbed problem to the unperturbed one $\boldsymbol{X}$, which is

$$\left\|\boldsymbol{Q}^*\widetilde{\boldsymbol{X}}\boldsymbol{Q} - \boldsymbol{X}\right\|$$

$$\leq \left\|\boldsymbol{U}_{\uparrow}^{\dagger}\right\| \left[\left(1+\frac{\left\|\boldsymbol{U}_{\uparrow}^{\dagger}\right\|}{1-\left\|\boldsymbol{U}_{\uparrow}^{\dagger}\right\|\left\|\widetilde{\boldsymbol{U}}_{\uparrow}\boldsymbol{Q}-\boldsymbol{U}_{\uparrow}\right\|}\right)\left\|\widetilde{\boldsymbol{U}}_{\uparrow}\boldsymbol{Q}-\boldsymbol{U}_{\uparrow}\right\|\|\boldsymbol{U}_{\downarrow}\boldsymbol{Q}\|+\left\|\widetilde{\boldsymbol{U}}_{\downarrow}\boldsymbol{Q}-\boldsymbol{U}_{\downarrow}\right\|\right].$$

Applying Lemma 2.1.18 yields

$$\begin{aligned}\left\|\boldsymbol{Q}^{*}\widetilde{\boldsymbol{X}}\boldsymbol{Q}-\boldsymbol{X}\right\| &\leq \left\|\boldsymbol{U}_{\uparrow}^{\dagger}\right\| \left[\left(1+\frac{\left\|\boldsymbol{U}_{\uparrow}^{\dagger}\right\|}{1-\left\|\boldsymbol{U}_{\uparrow}^{\dagger}\right\|\left\|\widetilde{\boldsymbol{U}}\boldsymbol{Q}-\boldsymbol{U}\right\|}\right)\left\|\widetilde{\boldsymbol{U}}\boldsymbol{Q}-\boldsymbol{U}\right\|\|\boldsymbol{U}\boldsymbol{Q}\|+\left\|\widetilde{\boldsymbol{U}}\boldsymbol{Q}-\boldsymbol{U}\right\|\right] \\ &\leq \left\|\boldsymbol{U}_{\uparrow}^{\dagger}\right\|\left\|\widetilde{\boldsymbol{U}}\boldsymbol{Q}-\boldsymbol{U}\right\| \left[\left(1+\frac{\left\|\boldsymbol{U}_{\uparrow}^{\dagger}\right\|}{1-\left\|\boldsymbol{U}_{\uparrow}^{\dagger}\right\|\left\|\widetilde{\boldsymbol{U}}\boldsymbol{Q}-\boldsymbol{U}\right\|}\right)\|\boldsymbol{U}\|+1\right] \\ &\leq \left\|\boldsymbol{U}_{\uparrow}^{\dagger}\right\|\left\|\widetilde{\boldsymbol{U}}\boldsymbol{Q}-\boldsymbol{U}\right\| \left(2+\frac{\left\|\boldsymbol{U}_{\uparrow}^{\dagger}\right\|}{1-\left\|\boldsymbol{U}_{\uparrow}^{\dagger}\right\|\left\|\widetilde{\boldsymbol{U}}\boldsymbol{Q}-\boldsymbol{U}\right\|}\right).\end{aligned}$$

Now, if we require the assumption $\|\boldsymbol{E}\| \leq \frac{1}{2}\sigma_M(\boldsymbol{H}_L)$ as we did in Lemma 4.3.2, then (C.0.1) provides

$$\left\|\widetilde{\boldsymbol{U}}\boldsymbol{Q}-\boldsymbol{U}\right\| \leq \frac{2\sqrt{2}\,\|\boldsymbol{E}\|_{\mathrm{F}}}{\sigma_M(\boldsymbol{H}_L)}$$

and therefore, together with (4.3.9) in the above nominator term

$$1-\left\|\boldsymbol{U}_{\uparrow}^{\dagger}\right\|\left\|\widetilde{\boldsymbol{U}}\boldsymbol{Q}-\boldsymbol{U}\right\| \geq 1-\frac{2\sqrt{2}\,\|\boldsymbol{E}\|_{\mathrm{F}}}{\alpha_{\min}\,\sigma_{\min}(\boldsymbol{A}_L)\,\sigma_{\min}(\boldsymbol{A}_{N-L+1})\,\sigma_{\min}(\boldsymbol{U}_{\uparrow})}.$$

We want to have this term to be greater or equal $\frac{1}{2}$ and thus, get the condition on the error matrix

$$\|\boldsymbol{E}\|_{\mathrm{F}} \leq \frac{\alpha_{\min}\,\sigma_{\min}(\boldsymbol{A}_L)\,\sigma_{\min}(\boldsymbol{A}_{N-L+1})\,\sigma_{\min}(\boldsymbol{U}_{\uparrow})}{4\sqrt{2}}.$$

This also implies the above assumption on $\|\boldsymbol{E}\|$. Altogether, since $\left\|\boldsymbol{U}_{\uparrow}^{\dagger}\right\| \geq 1$, we obtain

$$\begin{aligned}\left\|\boldsymbol{Q}^{*}\widetilde{\boldsymbol{X}}\boldsymbol{Q}-\boldsymbol{X}\right\| &\leq \left\|\boldsymbol{U}_{\uparrow}^{\dagger}\right\|\frac{2\sqrt{2}\,\|\boldsymbol{E}\|_{\mathrm{F}}}{\sigma_M(\boldsymbol{H}_L)}\left(2+2\left\|\boldsymbol{U}_{\uparrow}^{\dagger}\right\|\right) \leq 4\left\|\boldsymbol{U}_{\uparrow}^{\dagger}\right\|^{2}\frac{2\sqrt{2}\,\|\boldsymbol{E}\|_{\mathrm{F}}}{\sigma_M(\boldsymbol{H}_L)} \\ &\leq \frac{8\sqrt{2}\,\|\boldsymbol{E}\|_{\mathrm{F}}}{\alpha_{\min}\,\sigma_{\min}(\boldsymbol{A}_L)\,\sigma_{\min}(\boldsymbol{A}_{N-L+1})\,\sigma_{\min}(\boldsymbol{U}_{\uparrow})^{2}}.\end{aligned}$$

Finally, plugging this into (C.0.2) and estimating the constant yields

$$\mathrm{md}_{\mathbb{T}}(\Omega,\widetilde{\Omega}) \leq \frac{6(2M-1)\,\|\boldsymbol{A}_L\|\,\|\boldsymbol{E}\|_{\mathrm{F}}}{\alpha_{\min}\,\sigma_{\min}(\boldsymbol{A}_L)^{2}\,\sigma_{\min}(\boldsymbol{A}_{N-L+1})\,\sigma_{\min}(\boldsymbol{U}_{\uparrow})^{2}}.$$

Glossary of symbols

$\mathbb{N}$	natural numbers
$\mathbb{N}_+$	positive natural numbers
$\mathbb{Z}$	integers
$\mathbb{R}$	real numbers
$\mathbb{C}$	complex numbers
$\mathbb{T}$	periodic unit interval, $\mathbb{R}/\mathbb{Z} = [0, 1)$
$\mathrm{Re}(z)$	real part of z
$\mathrm{Im}(z)$	imaginary part z
$\arg(z)$	argument in $[0, 2\pi)$ of the complex number z
d	spatial dimension
$\boldsymbol{\nu}$	multi-index for degrees in $\mathbb{Z}^d$
N^d	number of columns of a Vandermonde matrix
M	number of nodes
S	number of clusters
λ	number of nodes in biggest cluster
L	ESPRIT parameter
Ω	set of nodes in $\mathbb{T}$ or $\mathbb{T}^d$
$\boldsymbol{A}, \boldsymbol{A}_N$	Vandermonde matrix of degree $N - 1$
$\boldsymbol{A}_L$	Vandermonde matrix of degree $L - 1$
$\boldsymbol{H}_L$	Hankel matrix with respect to L
$\mathbb{1}_{m\times n}$	matrix with each entry equal to one in $\mathbb{C}^{m\times n}$
$(\cdot)^\top$	transpose of a vector or a matrix
$\overline{(\cdot)}$	conjugate of a scalar, a vector or a matrix
$(\cdot)^*$	conjugate transpose of a vector or a matrix
$\mathrm{diag}(\cdot)$	diagonal matrix with given arguments on the main diagonal
$\boldsymbol{I}_m$	identity matrix in $\mathbb{C}^{m\times m}$
$\boldsymbol{0}_m$	matrix of zeros in $\mathbb{C}^{m\times m}$
$\boldsymbol{0}_{m\times n}$	matrix of zeros in $\mathbb{C}^{m\times n}$
dim	dimension of the space applied to
$\mathrm{range}(\boldsymbol{M})$	range/column space of a matrix $\boldsymbol{M}$
$\ker(\boldsymbol{M})$	kernel of a matrix $\boldsymbol{M}$

θ, θ_j	largest and j-th principal angle
cond	spectral condition number
$\sigma_{\min}$	smallest singular value
$\sigma_{\max}$	largest singular value
$\lambda_{\min}$	smallest eigenvalue of a Hermitian matrix
$\lambda_{\max}$	largest eigenvalue of a Hermitian matrix
$\langle \cdot, \cdot \rangle$	Euclidean scalar product
$\|\cdot\|$	Euclidean norm for vectors and spectral norm for matrices
$\|\cdot\|_{\mathrm{F}}$	Frobenius matrix norm
$\|\cdot\|_{\infty}$	max-norm for vectors and row-sum norm for matrices
$\|\cdot\|_1$	1-norm for vectors and column-sum norm for matrices
$\mathcal{C}(\mathbb{T}^d)$	set of continuous functions defined on $\mathbb{T}^d$
$L^1(\mathbb{T}^d)$	Lebesgue space of absolute integrable functions defined on $\mathbb{T}^d$
$L^2(\mathbb{T}^d)$	Lebesgue space of square integrable functions defined on $\mathbb{T}^d$
$\mathcal{M}(\mathbb{T}^d)$	Borel measures defined on $\mathbb{T}^d$
$\mathcal{P}(n)$	trigonometric polynomials of max-degree at most n
$\|\cdot\|_{\mathcal{C}(\mathcal{S})}$	max-norm for continuous functions defined on $\mathcal{S}$
$\|\cdot\|_{L^1(\mathbb{T}^d)}$	Lebesgue integral norm
$\|\cdot\|_{L^2(\mathbb{T}^d)}$	Lebesgue square integral norm
$\langle \cdot, \cdot \rangle_{L^2(\mathbb{T}^d)}$	scalar product on $L^2(\mathbb{T}^d)$
Γ	Gamma function
ζ	Riemann Zeta function
B	Beurling function
δ	Dirac measure
D_n	Dirichlet kernel of degree n
d_m	modified Dirichlet kernel of degree m
$\widehat{f}(k)$	k-th Fourier coefficient of the function f
$\widehat{\mu}(k)$	k-th Fourier coefficient of the measure μ
$\lceil x \rceil$	smallest integer larger or equal $x \in \mathbb{R}$
$\lfloor x \rfloor$	largest integer smaller or equal $x \in \mathbb{R}$
$[x]$	integer closest to $x \in \mathbb{R}$, taking the larger one if not unique
$\gtrsim$	greater or equal, up to a constant independent of quantities on the right hand side
$\lesssim$	less or equal, up to a constant independent of quantities on the right hand side

Bibliography

[1] E. Abbe. Beiträge zur Theorie des Mikroskops und der mikroskopischen Wahrnehmung. *Archiv f. mikrosk. Anatomie*, 9(1):413–468, 1873.

[2] A. Akinshin, D. Batenkov, and Y. Yomdin. Accuracy of spike-train Fourier reconstruction for colliding nodes. In *2015 International Conference on Sampling Theory and Applications (SampTA)*, pages 617–621, 2015.

[3] A. Akinshin, G. Goldman, V. Golubyatnikov, and Y. Yomdin. Accuracy of reconstruction of spike-trains with two near-colliding nodes. In *Complex analysis and dynamical systems VII*, volume 699 of *Contemp. Math.*, pages 1–17. Amer. Math. Soc., Providence, RI, 2017.

[4] G. E. Andrews, R. Askey, and R. Roy. *Special Functions*. Encyclopedia of Mathematics and its Applications. Cambridge University Press, 1999.

[5] C. Aubel. *Performance of super-resolution methods in parameter estimation and system identification*. PhD thesis, ETH Zürich, 2019.

[6] C. Aubel and H. Bölcskei. Deterministic performance analysis of subspace methods for cisoid parameter estimation. In *IEEE International Symposium on Information Theory, ISIT 2016, Barcelona, Spain, July 10-15, 2016*, pages 1551–1555. IEEE, 2016.

[7] C. Aubel and H. Bölcskei. Vandermonde matrices with nodes in the unit disk and the large sieve. *Appl. Comput. Harmon. Anal.*, 47(1):53–86, 2019.

[8] A. H. Barnett. How exponentially ill-conditioned are contiguous submatrices of the Fourier matrix?, Version 3. *arXiv e-prints*, Aug. 2020.

[9] D. Batenkov. Accurate solution of near-colliding Prony systems via decimation and homotopy continuation. *Theoret. Comput. Sci.*, 681:27–40, 2017.

[10] D. Batenkov. Stability and super-resolution of generalized spike recovery. *Appl. Comput. Harmon. Anal.*, 45(2):299–323, 2018.

[11] D. Batenkov, L. Demanet, G. Goldman, and Y. Yomdin. Conditioning of partial nonuniform Fourier matrices with clustered nodes. *SIAM J. Matrix Anal. Appl.*, 41(1):199–220, 2020.

[12] D. Batenkov, B. Diederichs, G. Goldman, and Y. Yomdin. The spectral properties of Vandermonde matrices with clustered nodes. *Linear Algebra Appl.*, 609:37–72, 2021.

[13] D. Batenkov, G. Goldman, and Y. Yomdin. Super-resolution of near-colliding point sources. *Information and Inference: A Journal of the IMA*, 05 2020. iaaa005.

[14] D. Batenkov and Y. Yomdin. Geometry and singularities of the Prony mapping. *J. Singul.*, 10:1–25, 2014.

[15] F. S. V. Bazán. Conditioning of rectangular Vandermonde matrices with nodes in the unit disk. *SIAM J. Matrix Anal. Appl.*, 21(2):679–693, 1999.

[16] F. Bazán. Sensitivity eigenanalysis for single shift-invariant subspace-based methods. *Signal Processing*, 80(1):89–100, 2000.

[17] B. Beckermann. The condition number of real Vandermonde, Krylov and positive definite Hankel matrices. *Numer. Math.*, 85(4):553–577, 2000.

[18] L. Berman and A. Feuer. On perfect conditioning of Vandermonde matrices on the unit circle. *Electron. J. Linear Algebra*, 16:157–161, 2007.

[19] A. Beurling. *The collected works of Arne Beurling. Vol. 2.* Contemporary Mathematicians. Birkhäuser Boston, Inc., Boston, MA, 1989. Harmonic analysis, Edited by L. Carleson, P. Malliavin, J. Neuberger and J. Wermer.

[20] G. Beylkin and L. Monzón. Approximation by exponential sums revisited. *Appl. Comput. Harmon. Anal.*, 28(2):131–149, 2010.

[21] Å. Björck and G. H. Golub. Numerical methods for computing angles between linear subspaces. *Math. Comp.*, 27(123):579–594, Jul. 1973.

[22] T. Blu, P. Dragotti, M. Vetterli, P. Marziliano, and L. Coulot. Sparse sampling of signal innovations. *IEEE Signal Processing Magazine*, 25(2):31–40, 2008.

[23] E. J. Candès and C. Fernandez-Granda. Towards a mathematical theory of super-resolution. *Comm. Pure Appl. Math.*, 67(6):906–956, 2014.

[24] S. Chen and A. Moitra. Algorithmic foundations for the diffraction limit. *arXiv e-prints*, Apr. 2020.

[25] M. Cheney. A mathematical tutorial on synthetic aperture radar. *SIAM Review*, 43(2):301–312, 2001.

[26] S. Chrétien and H. Tyagi. Multi-kernel unmixing and super-resolution using the modified matrix pencil method. *J. Fourier Anal. Appl.*, 26(1):Paper No. 18, 2020.

[27] A. Córdova Yévenes, W. Gautschi, and S. Ruscheweyh. Vandermonde matrices on the circle: spectral properties and conditioning. *Numer. Math.*, 57(6-7):577–591, 1990.

[28] A. Cuyt and W.-s. Lee. Multivariate exponential analysis from the minimal number of samples. *Adv. Comput. Math.*, 44(4):987–1002, 2018.

[29] A. P. de Camargo. An exponential lower bound for the condition number of real Vandermonde matrices. *Appl. Numer. Math.*, 128:81–83, 2018.

[30] B. G. R. de Prony. Essai éxperimental et analytique: sur les lois de la dilatabilité de fluides élastique et sur celles de la force expansive de la vapeur de l'alkool, a différentes températures. *Journal de l'école polytechnique*, 1(22):24–76, 1795.

[31] L. Demanet and N. Nguyen. The recoverability limit for superresolution via sparsity. *arXiv e-prints*, Feb. 2015.

[32] Q. Denoyelle, V. Duval, and G. Peyré. Support recovery for sparse super-resolution of positive measures. *J. Fourier Anal. Appl.*, 23(5):1153–1194, 2017.

[33] Q. Denoyelle, V. Duval, G. Peyré, and E. Soubies. The sliding Frank-Wolfe algorithm and its application to super-resolution microscopy. *Inverse Problems*, 36(1):014001, 42, 2020.

[34] B. Diederichs. Well-posedness of sparse frequency estimation. *arXiv e-prints*, May 2019.

[35] D. L. Donoho. Superresolution via sparsity constraints. *SIAM J. Math. Anal.*, 23(5):1309–1331, 1992.

[36] V. Duval and G. Peyré. Exact support recovery for sparse spikes deconvolution. *Found. Comput. Math.*, 15(5):1315–1355, 2015.

[37] M. Ehler, S. Kunis, T. Peter, and C. Richter. A randomized multivariate matrix pencil method for superresolution microscopy. *Electron. Trans. Numer. Anal.*, 51:63–74, 2019.

[38] P. Erdös and P. Turán. On interpolation. III. Interpolatory theory of polynomials. *Ann. of Math. (2)*, 41:510–553, 1940.

[39] A. Eriksson, P. Stoica, and T. Soderstrom. Second-order properties of MUSIC and ESPRIT estimates of sinusoidal frequencies in high SNR scenarios. *IEE Proceedings F - Radar and Signal Processing*, 140(4):266–272, 1993.

[40] D. G. Feingold and R. S. Varga. Block diagonally dominant matrices and generalizations of the Gerschgorin circle theorem. *Pacific J. Math.*, 12(4):1241–1250, 1962.

[41] P. J. S. G. Ferreira. Superresolution, the recovery of missing samples, and Vandermonde matrices on the unit circle. *Proceedings of the 1999 Workshop on Sampling Theory and Applications, Leon, Norway*, pages 216–220, 1999.

[42] W. Gautschi. Some elementary inequalities relating to the Gamma and incomplete Gamma function. *Journal of Mathematics and Physics*, 38(1-4):77–81, 1959.

[43] W. Gautschi. On inverses of Vandermonde and confluent Vandermonde matrices. *Numer. Math.*, 4:117–123, 1962.

[44] W. Gautschi. Optimally conditioned Vandermonde matrices. *Numer. Math.*, 24:1–12, 1975.

[45] W. Gautschi. On inverses of Vandermonde and confluent Vandermonde matrices. III. *Numer. Math.*, 29(4):445–450, 1977/78.

[46] W. Gautschi. How (un)stable are Vandermonde systems? In *Asymptotic and computational analysis (Winnipeg, MB, 1989)*, volume 124 of *Lecture Notes in Pure and Appl. Math.*, pages 193–210. Dekker, New York, 1990.

[47] W. Gautschi. *Walter Gautschi. Selected works with commentaries. Vol. 1.* Contemporary Mathematicians. Birkhäuser/Springer, New York, 2014. Edited by Claude Brezinski and Ahmed Sameh.

[48] W. Gautschi and G. Inglese. Lower bounds for the condition number of Vandermonde matrices. *Numer. Math.*, 52(3):241–250, 1988.

[49] G. H. Golub and C. F. Van Loan. *Matrix Computations.* Johns Hopkins Studies in the Mathematical Sciences. Johns Hopkins University Press, 3rd edition, 1996.

[50] K. Gröchenig. Reconstruction algorithms in irregular sampling. *Math. Comp.*, 59(199):181–194, 1992.

[51] K. Gröchenig. *Foundations of time-frequency analysis.* Applied and Numerical Harmonic Analysis. Birkhäuser Boston, Inc., Boston, MA, 2001.

[52] P. C. Hansen. The truncated SVD as a method for regularization. *BIT*, 27(4):534–553, 1987.

[53] N. J. Higham. *Accuracy and Stability of Numerical Algorithms.* Society for Industrial and Applied Mathematics, Philadelphia, PA, USA, second edition, 2002.

[54] R. A. Horn and C. R. Johnson. *Topics in Matrix Analysis.* Cambridge University Press, New York, USA, 1991.

[55] R. A. Horn and C. R. Johnson. *Matrix Analysis.* Cambridge University Press, New York, USA, 2nd edition, 2013.

[56] Y. Hua and T. K. Sarkar. Matrix pencil method for estimating parameters of exponentially damped/undamped sinusoids in noise. *IEEE Trans. Acoust. Speech Signal Process.*, 38(5):814–824, 1990.

[57] A. E. Ingham. Some trigonometrical inequalities with applications to the theory of series. *Math. Z.*, 41(1):367–379, 1936.

[58] I. C. F. Ipsen. *Numerical Matrix Analysis.* Society for Industrial and Applied Mathematics, 2009.

[59] L. Kämmerer, S. Kunis, and D. Potts. Interpolation lattices for hyperbolic cross trigonometric polynomials. *J. Complexity*, 28(1):76–92, 2012.

[60] D. W. Kammler. *A First Course in Fourier Analysis.* Cambridge University Press, 2nd edition, 2007.

[61] I. Keller and G. Plonka. Modifications of Prony's method for the recovery and sparse approximation of generalized exponential sums. *arXiv e-prints*, Jan. 2020.

[62] A. Klinger. The Vandermonde matrix. *Amer. Math. Monthly*, 74:571–574, 1967.

[63] V. Komornik and P. Loreti. *Fourier Series in Control Theory.* Springer-Verlag, New York, 2005.

[64] V. Komornik and P. Loreti. Semi-discrete Ingham-type inequalities. *Appl. Math. Optim.*, 55(2):203–218, 2007.

[65] H. Krim and M. Viberg. Two decades of array signal processing research: the parametric approach. *IEEE Signal Processing Magazine*, 13(4):67–94, 1996.

[66] S. Kunis, H. M. Möller, T. Peter, and U. von der Ohe. Prony's method under an almost sharp multivariate Ingham inequality. *J. Fourier Anal. Appl.*, 24(5):1306–1318, 2018.

[67] S. Kunis, H. M. Möller, and U. von der Ohe. Prony's method on the sphere. *SMAI J. Comput. Math.*, S5:87–97, 2019.

[68] S. Kunis and D. Nagel. On the condition number of Vandermonde matrices with pairs of nearly-colliding nodes. *Numer. Algorithms*, Jul. 2020.

[69] S. Kunis and D. Nagel. On the smallest singular value of multivariate Vandermonde matrices with clustered nodes. *Linear Algebra Appl.*, 604:1–20, 2020.

[70] S. Kunis, T. Peter, T. Römer, and U. von der Ohe. A multivariate generalization of Prony's method. *Linear Algebra Appl.*, 490:31–47, 2016.

[71] S. Kunis and D. Potts. Stability results for scattered data interpolation by trigonometric polynomials. *SIAM J. Sci. Comput.*, 29(4):1403–1419, 2007.

[72] W. Li and W. Liao. Conditioning of restricted Fourier matrices and super-resolution of MUSIC. In *2019 13th International conference on Sampling Theory and Applications (SampTA)*, pages 1–4, 2019.

[73] W. Li and W. Liao. Stable super-resolution limit and smallest singular value of restricted Fourier matrices. *Appl. Comput. Harmon. Anal.*, 51:118–156, 2021.

[74] W. Li, W. Liao, and A. Fannjiang. Super-resolution limit of the ESPRIT algorithm. *IEEE Trans. Inform. Theory*, 66(7):4593–4608, 2020.

[75] W. Liao. MUSIC for multidimensional spectral estimation: stability and super-resolution. *IEEE Trans. Signal Process.*, 63(23):6395–6406, 2015.

[76] W. Liao and A. Fannjiang. MUSIC for single-snapshot spectral estimation: stability and super-resolution. *Appl. Comput. Harmon. Anal.*, 40(1):33–67, 2016.

[77] F. R. S. Lord Rayleigh. Xxxi. Investigations in optics, with special reference to the spectroscope. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, 8(49):261–274, 1879.

[78] N. Macon and A. Spitzbart. Inverses of Vandermonde matrices. *Amer. Math. Monthly*, 65:95–100, 1958.

[79] L. Mattner and B. Roos. Maximal probabilities of convolution powers of discrete uniform distributions. *Statist. Probab. Lett.*, 78(17):2992–2996, 2008.

[80] A. Moitra. Super-resolution, extremal functions and the condition number of Vandermonde matrices. In *STOC'15—Proceedings of the 2015 ACM Symposium on Theory of Computing*, pages 821–830. ACM, New York, 2015.

[81] H. L. Montgomery and R. C. Vaughan. Hilbert's inequality. *J. London Math. Soc. (2)*, 8:73–82, 1974.

[82] V. I. Morgenshtern. Super-resolution of positive sources on an arbitrarily fine grid. *arXiv e-prints*, May 2020.

[83] V. I. Morgenshtern and E. J. Candès. Super-resolution of positive sources: the discrete setup. *SIAM J. Imaging Sci.*, 9(1):412–444, 2016.

[84] M. Negreanu and E. Zuazua. Discrete Ingham inequalities and applications. *SIAM J. Numer. Anal.*, 44(1):412–448, 2006.

[85] V. Y. Pan. How bad are Vandermonde matrices? *SIAM J. Matrix Anal. Appl.*, 37(2):676–694, 2016.

[86] V. Pereyra and G. Scherer. Exponential data fitting. *Exponential Data Fitting and its Application, edited by V. Pereyra and G. Scherer (Bentham Sci. Publ.)*, pages 1–26, 2010.

[87] T. Peter and G. Plonka. A generalized Prony method for reconstruction of sparse sums of eigenfunctions of linear operators. *Inverse Problems*, 29(2):025001, 21, 2013.

[88] V. F. Pisarenko. The retrieval of harmonics from a covariance function. *Geophysical Journal of the Royal Astronomical Society*, 33(3):347–366, 1973.

[89] G. Plonka, D. Potts, G. Steidl, and M. Tasche. *Numerical Fourier analysis*. Applied and Numerical Harmonic Analysis. Birkhäuser/Springer, Cham, 2018.

[90] G. Plonka and M. Tasche. Prony methods for recovery of structured functions. *GAMM-Mitt.*, 37(2):239–258, 2014.

[91] D. Potts and M. Tasche. Parameter estimation for exponential sums by approximate Prony method. *Signal Processing*, 90(5):1631–1642, 2010.

[92] D. Potts and M. Tasche. Parameter estimation for multivariate exponential sums. *Electron. Trans. Numer. Anal.*, 40:204–224, 2013.

[93] D. Potts and M. Tasche. Parameter estimation for nonincreasing exponential sums by Prony-like methods. *Linear Algebra Appl.*, 439(4):1024–1039, 2013.

[94] D. Potts and M. Tasche. Error estimates for the ESPRIT algorithm. In *Large truncated Toeplitz matrices, Toeplitz operators, and related topics*, volume 259 of *Oper. Theory Adv. Appl.*, pages 621–648. Birkhäuser/Springer, Cham, 2017.

[95] H. Robbins. A remark on Stirling's formula. *Amer. Math. Monthly*, 62(1):26–29, 1955.

[96] R. Roy and T. Kailath. ESPRIT - estimation of signal parameters via rotational invariance techniques. *IEEE Trans. Acoust. Speech Signal Process.*, 37(7):984–995, 1989.

[97] W. Rudin. *Real and complex analysis*. McGraw-Hill Book Co., New York, third edition, 1987.

[98] T. Sauer. Prony's method in several variables. *Numer. Math.*, 136(2):411–438, 2017.

[99] R. O. Schmidt. Multiple emitter location and signal parameter estimation. *IEEE Trans. Antennas and Propagation*, 34(3):276–280, 1986.

[100] A. Selberg. *Collected papers. Vol. II*. Springer-Verlag, Berlin, 1991. With a foreword by K. Chandrasekharan.

[101] S. Serra-Capizzano. An elementary proof of the exponential conditioning of real Vandermonde matrices. *Boll. Unione Mat. Ital. Sez. B Artic. Ric. Mat. (8)*, 10(3):761–768, 2007.

[102] J. M. Steele. *The Cauchy-Schwarz Master Class: An Introduction to the Art of Mathematical Inequalities*. Cambridge University Press, 2004.

[103] G. W. Stewart. *Matrix Algorithms: Volume 1: Basic Decompositions*. Other Titles in Applied Mathematics. Society for Industrial and Applied Mathematics, 1998.

[104] G. W. Stewart and J.-G. Sun. *Matrix Perturbation Theory*. Computer science and scientific computing. Academic Press, 1990.

[105] P. Stoica and R. Moses. *Spectral Analysis of Signals*. Pearson Prentice Hall, 2005.

[106] P. Stoica and T. Soderstrom. Statistical analysis of MUSIC and subspace rotation estimates of sinusoidal frequencies. *IEEE Trans. Signal Process.*, 39(8):1836–1847, 1991.

[107] A. Strotmann. *Konstruktion einer in Ort und Frequenz lokalisierten Funktion*. Bachelor thesis, Osnabrück University, 2020.

[108] G. Tang, B. N. Bhaskar, P. Shah, and B. Recht. Compressed sensing off the grid. *IEEE Trans. Inform. Theory*, 59(11):7465–7490, 2013.

[109] E. E. Tyrtyshnikov. How bad are Hankel matrices? *Numer. Math.*, 67(2):261–269, 1994.

[110] J. D. Vaaler. Some extremal functions in Fourier analysis. *Bull. Amer. Math. Soc. (N.S.)*, 12(2):183–216, 1985.

[111] M. Vetterli, P. Marziliano, and T. Blu. Sampling signals with finite rate of innovation. *IEEE Trans. Signal Process.*, 50(6):1417–1428, 2002.

[112] P.-Å. Wedin. Perturbation bounds in connection with singular value decomposition. *BIT Numerical Mathematics*, 12(1), Mar. 1972.

[113] P.-Å. Wedin. Perturbation theory for pseudo-inverses. *BIT Numerical Mathematics*, 13(2), Jun. 1973.

[114] J. G. Wendel. Note on the Gamma function. *Amer. Math. Monthly*, 55(9):563–564, 1948.

[115] E. T. Whittaker and G. N. Watson. *A Course of Modern Analysis*. Cambridge Mathematical Library. Cambridge University Press, 4th edition, 1996.

[116] Y. Yu, T. Wang, and R. J. Samworth. A useful variant of the Davis-Kahan theorem for statisticians. *Biometrika*, 102(2):315–323, 2015.

Curriculum vitae

Personal information

Name	Dominik Nagel
Date of Birth	14 August 1991
Place of Birth	Ochtrup
Nationality	German
Address	Mozartstraße 1
	49078 Osnabrück
E-mail	dnagel@uos.de

Education

since 08/2016	Research assistant, Institute for Mathematics, Osnabrück University
2013 – 2015	Master studies, RWTH Aachen University
2010 – 2013	Bachelor studies, RWTH Aachen University
1998 – 2010	School, Abitur