Universidad
Carlos III de Madrid

# TESIS DOCTORAL

## STRUCTURED CONDITION NUMBERS FOR PARAMETERIZED QUASISEPARABLE MATRICES


Autor:
Kenet Jorge Pomés Portal

Director:
Froilán Martínez Dopico

Departamento de matemáticas
Universidad Carlos III de Madrid


Leganés, Septiembre, 2016

Universidad
Carlos III de Madrid

TESIS DOCTORAL

# STRUCTURED CONDITION NUMBERS FOR PARAMETERIZED QUASISEPARABLE MATRICES


Autor:

Kenet Jorge Pomés Portal

Director:

Froilán Martínez Dopico


Firma del Tribunal Calificador:

| | | Firma |
|---|---|---|
| Presidente: | _____ | _____ |
| Vocal: | _____ | _____ |
| Secretario: | _____ | _____ |

Calificación:

Leganés, de de 201

Este trabajo ha sido desarrollado en el Departamento de Matemáticas de la Universidad Carlos III de Madrid (UC3M) bajo la dirección del profesor Froilán Martínez Dopico. Se contó con una beca de la UC3M de ayuda al estudio de máster y posteriormente con un contrato predoctoral de la UC3M. Adicionalmente se recibió ayuda parcial de los proyectos de investigación: "Structured Numerical Linear Algebra: Matrix Polynomials, Special Matrices, and Conditioning" (Ministerio de Economía y Competitividad de España, Número de proyecto: MTM2012-32542) y "Structured Numerical Linear Algebra for Constant, Polynomial and Rational Matrices" (Ministerio de Economía y Competitividad de España, Número de proyecto: MTM2015-65798-P).

# Agradecimientos

En primer lugar, quiero agradecer sinceramente a mi director de tesis, Froilán Martínez Dopico, por todo su esfuerzo y por su apoyo en todo momento. Su impresionante paciencia y su rigor me han permitido llegar hasta aquí y me dejarán marcado como profesional y como persona. !Muchas gracias por aceptarme como alumno y por guiarme!

A Lula, mi amiga, novia y esposa, gracias por todo lo que no puedo escribir aquí.

A mis padres por su apoyo y amor desde siempre y ahora desde la distancia. A mi hermano que lucha conmigo día a día.

A Héctor, gracias a él estoy aquí. Gracias, sobre todo, por tu amistad.

A mis amigos del Despacho 2.1D18, Lino y Luis Miguel. A Javier Pérez que tanto me ayudó al principio y en tantas otras ocasiones. A Javier González, por todo su tiempo, esfuerzo y apoyo. A Fredy por ser la persona que es. A Andys y Amauris. A Walter y Yadira. A mis demás amigos, Pitu, Lolo, William, Julito, Sonia y Toñi.

Al Departamento de Matemáticas en general. A los profesores Julio Moro, Fernando de Terán, José Manuel Rodríguez, Paco Marcellán, Antonio García y Héctor Pijeira, en particular. Al imprescindible Alberto Calvo.

# Abstract

Low-rank structured matrices have attracted much attention in the last decades, since they arise in many applications and all share the fundamental property that can be represented by $\mathcal{O}(n)$ parameters, where $n \times n$ is the size of the matrix. This property has allowed the development of fast algorithms for solving numerically many problems involving low-rank structured matrices by performing operations on the parameters describing the matrices, instead of directly on the matrix entries. Among these problems the solution of linear systems of equations and the computation of the eigenvalues are probably the most basic and relevant ones. Therefore, it is important to measure, via structured computable condition numbers, the relative sensitivity of the solutions of linear systems with low-rank structured coefficient matrices, and of the eigenvalues of those matrices, with respect to relative perturbations of the parameters representing such matrices, since this sensitivity determines the maximum accuracy attainable by fast algorithms and allows us to decide which set of parameters is the most convenient from the point of view of accuracy.

In this PhD Thesis we develop and analyze condition numbers for eigenvalues of low-rank matrices and for the solutions of linear systems involving such matrices. To this purpose, general expressions are obtained for the condition numbers of the solution of a linear system of equations whose coefficient matrix is any differentiable function of a vector of parameters with respect to perturbations of such parameters, and also for the eigenvalues of those matrices.

Since there are many different classes of low-rank structured matrices and many different types of parameters describing them, it is not possible to cover all of them in this thesis. Therefore, the general expressions of the condition numbers are particularized to the important case of quasiseparable matrices and to the quasiseparable and the Givens-vector representations. In the case of $\{1,1\}$-quasiseparable matrices, we provide explicit expressions of the corresponding condition numbers for these two representations that can be estimated in $\mathcal{O}(n)$ operations. In addition, detailed theoretical and numerical comparisons of the condition numbers with respect to these two representations between themselves, and with respect to unstructured condition numbers are provided. These comparisons show that there are situations in which the unstructured condition numbers are much larger than the structured ones, but that the opposite never happens. On the other hand, for general $\{n_L, n_U\}$-quasiseparable matrices we also present an explicit expression for computing the eigenvalue condition number for the general quasiseparable representation in $\mathcal{O}((n_L^2 + n_U^2)n)$ operations. In this case, significant differences are obtained with respect to the $\{1,1\}$-case, since if $n_L$ or $n_U$ are greater than one, then the structured condition number may be much larger than the unstructured one, which suggests the use of a more stable representation in that case. The approach presented in

this dissertation can be generalized to other classes of low-rank structured matrices and parameterizations, as well as to any class of structured matrices that can be represented by parameters, independently of whether or not they enjoy a "low-rank" structure.

# Contents

# Chapter 1

# Introduction and summary of main results

In simple words, a low-rank structured matrix is a matrix such that large submatrices of it have ranks much smaller than the size of the matrix. Perhaps, the best known examples of low-rank structured matrices are tridiagonal and other banded matrices with small bandwidth, for which all the submatrices lying in the (strictly) lower or upper triangular parts have ranks smaller than or equal to the bandwidth. These examples correspond to special cases of sparse matrices, but many other classes of *dense* low-rank structured matrices are available in the literature and arise in many applications. Research on low-rank structured matrices has received much attention in the last 15 years from the points of view of theory, computations, and applications. In fact, a number of recent books are devoted to this subject [27, 28, 61, 62], as well as survey papers [13], and the interested reader can find a huge number of references on this topic in them. From a numerical perspective, the key features of $n \times n$ low-rank structured matrices are that they can be very often described in terms of different sets of $O(n)$ parameters, called *representations* [61, Ch. 2], and that this fact has been used to develop many *fast algorithms* operating on these parameters to perform computations with low-rank structured matrices [27, 28, 61, 62]. In this context, fast algorithms mean algorithms with cost $O(n)$ operations for solving linear systems of equations or with cost $O(n^2)$ operations for solving eigenvalue problems, which should be compared with the $O(n^3)$ cost of traditional dense matrix algorithms [38, 41].

Besides being the subject of modern research, low-rank structured matrices have an old and long history. One of the first examples of low-rank structured matrices are the *single-pair* matrices presented in 1941 in [36] in the context of totally non-negative matrices (see also [35]). Another historical source of low-rank structured matrices is related to the efforts made in the 1950s to compute inverses of tridiagonal and, in general, of banded matrices with small bandwidth [1, 2, 7, 55]. These efforts were motivated by early research on the numerical solution of certain integral

equations, boundary value problems, and problems in statistics. Inverses of banded matrices are included in a class of low-rank structured matrices called nowadays *semiseparable* matrices [61, Theorems 1.38 and 8.45]. Since the 1950s, the number of publications on low-rank structured matrices has increased considerably and, in fact, has exploded in the last 15 years. We refer the reader to the historical notes in [27, 28, 61, 62] and the detailed bibliography in [60].

Many interesting applications of low-rank structured matrices are discussed in the general references [13, 27, 28, 61, 62], but here we would like to emphasize a few of them and to cite a few specific references as a sample. Fast computations with low-rank structured matrices have been used, for instance, in the numerical solution of elliptic partial differential equations [5, 39], in the numerical solution of integral equations [12, 50, 51], and in the classical problem of computing all the roots of a polynomial of degree $n$ via matrix eigenvalue algorithms with cost of $O(n^2)$ operations and $O(n)$ storage [9, 11, 14, 29, 58]. With respect to this last problem, the recent reference [3] deserves special attention, since it includes a new algorithm and, for the first time in the literature, a rigorous proof that a fast and memory efficient algorithm for computing all the roots of a polynomial is backward stable in a matrix sense, which solves a long-standing open problem in Numerical Linear Algebra.

An important drawback of fast algorithms for low-rank structured matrices is that they have not been proved to be backward stable, with the exception of the particular ones in [3, 6, 21, 57, 66, 67]. Taking into account the large number of references available on these algorithms, this lack of error analyses is striking. Possible reasons for it are that these fast algorithms are often involved, which makes the potential errors analyses very difficult and, also, that some of them are potentially unstable in rare cases. In this scenario, a practical option is to estimate *a posteriori* error bounds for the outputs of these algorithms based on the classical approach in Numerical Linear Algebra of computing the residuals of the computed quantities, which give the backward errors, and multiply them by the corresponding condition numbers [38, 41, 42, 43]. Since fast algorithms for low-rank structured matrices operate on parameters and not on matrix entries, the most sensible approach would be to estimate from the residuals the backward errors in the parameters defining the matrix, and to multiply them by the corresponding condition numbers with respect to perturbations of those parameters. The results in this thesis include the first steps in this ambitious plan, since we present condition numbers with respect to parameters for a family of low-rank structured matrices. More precisely, we develop condition numbers for eigenvalues and for the solution of linear systems of equations, and show that for the case of $\{1,1\}$-quasiseparable matrices some eigenvalues or solution vectors may be extremely ill-conditioned under general componentwise relative unstructured perturbations of the matrix entries, but very well-conditioned under perturbations in the parameters.

There exist many classes of low-rank structured matrices and it is not possible to

cover all of them in this thesis. Therefore, we restrict ourselves to the particular but important class of quasiseparable matrices, whose definition is recalled in Section 2.2. This class of matrices was introduced in [24] and includes several other relevant classes of low-rank structured matrices, as is discussed in [61, p. 10]. In addition, we would like to emphasize that the approach presented in this work can be easily extended to other classes of structured matrices as long as they are explicitly described in terms of parameters, as a consequence of the general framework developed in Sections 3.1.2 and 3.2.2. So, we expect that the results in this thesis can show how to get in the future condition numbers for many other classes of low-rank structured matrices and problems, as well as to foster more research on this topic.

The results in Chapters 4, 5, and 6 can be seen as a new contribution to structured perturbation theory, a very fruitful and active area of research inside Numerical Linear Algebra. The general goal of the research in this area is to show that either for matrices in certain classes or for perturbations with particular properties, it is possible to derive much stronger perturbation bounds for the solutions of numerical problems (finding eigenvalues or solving linear systems, for instance) than the traditional ones obtained for general unstructured perturbations, and that these strong bounds can be used to prove that certain algorithms taking advantage of the structure yield much more accurate outputs than standard unstructured algorithms. The number of publications in this area is also very large and, here, we simply list a small sample of relevant references [32, 42, 43, 45, 46, 49]. A common thread in structured perturbation theory is that the relative, instead the absolute, sensitivity of the desired output of an algorithm is studied and bounded, as a consequence of the high expectations of the computations in our times. In addition, for the same reasons, many works on structured perturbation theory consider relative *componentwise* perturbations of the parameters defining the matrices of the given problem. We follow both approaches in this dissertation, which are also motivated by the fact that the parameters defining a given quasiseparable matrix can be widely scaled, while yielding the same matrix [61, Chs. 1 & 2], and so their collective norm is not related to the norm of the matrix. The results in this dissertation are, in particular, influenced by the recent ones in [32], but also influenced by the classical and seminal reference [53], which is often forgotten and which initiated the use of differential calculus for getting condition numbers.

Another goal of this dissertation is to provide a way to compare different representations of low-rank structured matrices. It is well known that the same quasiseparable matrix can be represented by different sets of parameters ([26], [61, Ch. 2], see also Section 2.3 in this thesis), also called generators, and it is not clear which set is more appropriate for developing a fast algorithm. A sensible option is to choose that representation for which the condition number of the desired quantity with respect to perturbations of the parameters is the smallest one. For this reason, we study and compare condition numbers for eigenvalues of $\{1,1\}$-quasiseparable matrices and for the solutions of linear systems of equations with a $\{1,1\}$-quasiseparable

coefficient matrix with respect to relative perturbations on the parameters in different representations of such quasiseparable matrices. More precisely, we consider all the quasiseparable representations [26], there are infinitely many, and the, essentially unique, Givens-vector representation ([59], [61, Ch. 2]), and we prove that, in both cases, the condition numbers for the different representations have similar magnitudes, but that those corresponding to the Givens-vector representation are the respective smallest ones. We advance that the most basic reason for this fact is the presence of extra constraints in the parameters of the Givens-vector representation with respect to the ones of the quasiseparable representation, which restrict the set of possible perturbations. In this context, it should be stressed that relative condition numbers do not take into account other issues which are also very important in practical computations, as the appearance of very large or small parameters that can produce overflow or underflow and spoil the whole computation.

Two remarkable unexpected properties are proved in this dissertation for the covered condition numbers for $\{1,1\}$-quasiseparable matrices with respect to the (infinitely many) quasiseparable representations. First, that these condition numbers are independent of the particular representation (see Proposition 4.4 for the solutions of linear systems and Proposition 5.3 for the eigenvalues case) and, second, that they can be respectively expressed just in terms of the matrix entries, i.e., without using any parametrization of the matrix (see Theorems 4.3 and 5.2, respectively). Nevertheless, the low-rank structure of the matrix is reflected in the way the different entries of the matrix contribute to the conditions number. These properties are important because it is not always trivial to compute a parametrization of a low-rank structured matrix.

On the other hand, for the general case of $\{n_L, n_U\}$-quasiseparable matrices, we also derived an expression for computing the eigenvalue condition number with respect to the general quasiseparable representation, but in this case, that condition number does depend on the parameters (see the expression in Theorem 6.2 for its computation), and, what it is more important, it can be much larger than the unstructured standard condition number as we have seen in our numerical experiments described in Section 6.3. This significant difference with the $\{1,1\}$-case suggests the need of studying different and more stable representations for generating those matrices and operating on them. We are currently working on this problem using the description of the Givens-weight representation recently presented in [57].

The rest of the dissertation is organized as follows. In Chapter 2 we introduce the notions of structures and rank structured matrices, define quasiseparable matrices using the notation in [61], and recall some of its basic properties. The quasiseparable and the Givens-vector representations for $\{1,1\}$-quasiseparable matrices are also presented in the way that they will be used trough the rest of the dissertation. In addition, for the general class of $\{n_L, n_U\}$-quasiseparable matrices, an extension of the quasiseparable representation is also provided. This chapter is based in the book [61], and does not include original results from the author.

Chapter 3 is organized in two different parts. The first one, Section 3.1, is devoted to the study of condition numbers for linear systems. In particular, in Section 3.1.1, the definitions of standard normwise and componentwise condition numbers for the solution of a linear system of equations are presented together with explicit expressions for their computations and some of their main properties. In Section 3.1.2, we introduce the notion of a structured condition number for the solution of a linear system of equations with a parameterized matrix of coefficients with respect to relative componentwise perturbations of the parameters defining the matrix, and we provide an explicit expression for its computation. The second part of the chapter, Section 3.2, is devoted to condition numbers for simple eigenvalues and is organized in a similar way to the previous one, by presenting the analogous standard normwise and componentwise unstructured eigenvalue condition numbers in Section 3.2.1 and the structured eigenvalue condition number for parameterized matrices in Section 3.2.2. The results in Sections 3.1.1 and 3.2.1 are not original contributions of the author of this thesis, and are based in the book by Higham [41], the reference [42] by Higham & Higham, and standard references as [38] for the classical Wilkinson eigenvalue condition number. On the other hand, the expressions for the condition numbers of parameterized matrices in Sections 3.1.2 and 3.2.2 are original contributions of the author that are used in further chapters for getting explicit expressions of the condition numbers for the representations of quasiseparable matrices covered in this dissertation. The original results in Sections 3.1.2 and 3.2.2 are respectively included in [23] and [22].

Chapter 4 is devoted to the conditioning of the solutions of linear systems of equations with $\{1,1\}$-quasiseparable matrices of coefficients with respect to relative componentwise perturbations of the parameters generating the matrix. In particular, in Section 4.1, we deduce an expression for computing the structured condition number for linear systems in the quasiseparable representation and we prove some of its more important properties, like its invariance under different sets of quasiseparable parameters when the tolerances for the relative perturbations are measured with respect to the set of parameters itself, its invariance under diagonal scaling of the coefficient matrix on the left, and more important, that the structured condition number can not be much greater than the unstructured one (but it can be much smaller as we observed in the numerical experiments described in Section 4.5). In Section 4.2, we define the Givens-vector representation via tangents of a $\{1,1\}$-quasiseparable matrix and obtain an expression for the computation of the structured condition number for linear systems in that representation. In Section 4.3, a comparison of the structured condition numbers under the quasiseparable and the Givens-vector representation via tangents is provided, proving that, in the $\{1,1\}$-case, the Givens-vector is a more stable representation since it produces condition numbers that are smaller than the ones for the quasiseparable representation. In Section 4.4, we show how to estimate fast the structured condition numbers via an effective condition number. In order to corroborate some of the results obtained

trough the chapter, a brief description of some numerical experiments we have performed is provided in Section 4.5. All the original results in this chapter are included in [23].

Chapter 5 follows a similar structure to that of Chapter 4, but, in this case, we consider eigenvalue condition numbers also for $\{1,1\}$-quasiseparable matrices. In Section 5.1, a formula for computing the structured eigenvalue condition number in the quasiseparable representation is deduced. We prove that this condition number is also invariant under different quasiseparable representations when the tolerances for the relative perturbations are measured against the set of parameters itself, and under diagonal similarities as well. Furthermore, it is proved that in the $\{1,1\}$-case the quasiseparable structure plays a key role in the accuracy of eigenvalue computations since the structured condition number can be much smaller than the unstructured one but the opposite can not happen. In Section 5.2, it is proved that this condition number can be computed fast via an algorithm with $\mathcal{O}(n)$ cost. In Section 5.3 we provide an expression for computing the eigenvalue condition number in the Givens-vector representation via tangents, and in Section 5.4, we provide an algorithm with $\mathcal{O}(n)$ cost for its computation. In Section 5.5, a comparison of the condition numbers in the quasiseparable and in the Givens-vector representations have been carried out in order to prove that the Givens-vector representation is also more stable for eigenvalue computations than the quasiseparable one since its produces smaller eigenvalue condition numbers. The chapter is concluded by describing some numerical experiments in Section 5.6. All the original contributions in this chapter are included in [22].

In Chapter 6, we start the study of the conditioning of simple eigenvalues for general $\{n_L, n_U\}$-quasiseparable matrices. In particular, in Section 6.1, we obtain an expression for computing the structured eigenvalue condition number in the general quasiseparable representation for an $\{n_L, n_U\}$-quasiseparable matrix. In this case, significant differences with the $\{1,1\}$-case are pointed out since when $n_L > 1$ or $n_U > 1$, the structured condition number does depend on the choice of the parameters and more important, the structured condition number can be much larger than the unstructured one. Therefore, we provide an interpretation of this fact and obtain a bound in terms of an unstructured condition number. In Section 6.2, it is proved that this structured condition number can be computed fast via an $\mathcal{O}((n_L^2 + n_U^2)n)$ algorithm, and in Section 6.3, a brief description of some numerical experiments is provided. Most of the results in this chapter are original contributions of the author, but have not been submitted yet for publication since we want to complete these results with the study of eigenvalue condition numbers in the Givens-weight representation for general $\{n_L, n_U\}$-quasiseparable matrices [17, 57]. We are currently working on this problem, which is hard since the Givens-weight representation of $\{n_L, n_U\}$-quasiseparable matrices is not easy to describe explicitly.

**Notation.** Given the field of scalars $\mathbb{F} \in \{\mathbb{R}, \mathbb{C}\}$, we denote by $\mathbb{F}^{m\times n}$ to the set of matrices of size $m \times n$ with entries in $\mathbb{F}$. We will follow a common notation

in Numerical Linear Algebra and use capital Roman letters $A$, $B$,..., for matrices, lower case Roman letters $\boldsymbol{x}, \boldsymbol{y}, \ldots$ for column vectors, and Greek letters $\alpha, \beta, \ldots$, for scalars. Except in the preliminary Section 3.2, only real matrices are considered, but some eigenvalues and eigenvectors may be complex. We will denote by $\boldsymbol{e}_i$ to the $i$-th vector in the canonical basis of $\mathbb{R}^n$, i.e., $\boldsymbol{e}_i$ denotes the $i$-th column of the identity matrix $I_n$ in $\mathbb{R}^{n\times n}$. Given a complex column vector $\boldsymbol{y}$ of size $n \times 1$, $\boldsymbol{y}^T$ denotes its transpose, and $\boldsymbol{y}^* := (\overline{\boldsymbol{y}})^T$ its conjugate transpose, where $\overline{\alpha}$ is the conjugate of $\alpha$ and conjugation of vectors should be understood in a componentwise sense. We consider the following usual norms

$$\|\boldsymbol{y}\|_1 := \sum_{i=1}^{n} |y_i|\,, \quad \|\boldsymbol{y}\|_2 := \left(\sum_{i=1}^{n} |y_i|^2\right)^{\frac{1}{2}}, \quad \text{and} \quad \|\boldsymbol{y}\|_\infty := \max_{1\le i\le n} |y_i|,$$

where $y_i$ denotes the $i$-th coordinate of the vector $\boldsymbol{y}$, and the corresponding operator norms for $m \times n$ matrices [38, 41]:

$$\|A\|_1 = \max_{1\le j\le n} \sum_{i=1}^{m} |a_{ij}|\,, \;\; \|A\|_\infty = \max_{1\le i\le n} \sum_{j=1}^{n} |a_{ij}|\,, \;\; \text{and} \;\; \|A\|_2 = \left(\rho\left(A^*A\right)\right)^{\frac{1}{2}} = \sigma_{\max}(A),$$

where, for any square matrix $M$, $\rho(M)$ denotes the spectral radius of $M$, i.e.:

$$\rho(M) = \max\{|\lambda| : \lambda \,\text{is an eigenvalue of}\, M\},$$

and $\sigma_{\max}(A)$ denotes the largest singular value of $A$.

For any square matrix $M$, we write the eigenvalue-eigenvector equations as $M\boldsymbol{x} = \lambda\boldsymbol{x}$ and $\boldsymbol{y}^*M = \lambda\boldsymbol{y}^*$, where $\boldsymbol{y}$ and $\boldsymbol{x}$ denote, respectively, the left and right eigenvectors associated to the eigenvalue $\lambda$ of $M$.

Standard MATLAB notation for submatrices is used, i.e., given a matrix $M \in \mathbb{R}^{m\times n}$, the expression $M(i:j, k:l)$, where $1 \le i \le j \le m$ and $1 \le k \le l \le n$, denotes the submatrix of $M$ consisting of the intersection of rows $i$ up to and including $j$ of $M$ and of columns $k$ up to and including $l$ of $A$.

# Chapter 2

# A brief review of quasiseparable matrices

As its name suggests, this chapter is devoted to define quasiseparable matrices which is the main class of matrices covered in this thesis, and to recall some of their basic properties. In Section 2.1 we recall some concepts related to quasiseparable matrices like the definitions of low-rank structured matrices, lower and upper triangular structures, and the subclass of semiseparable matrices. In Section 2.2, quasiseparable matrices are defined and some of their representations are introduced in Section 2.3. This chapter is mainly based on Chapters 1, 2 and 8 from the book by Vandebril, Van Barel and Mastronardi [61].

## 2.1 Basics on low-rank structured matrices

A rank structured matrix is, in plain words, a matrix such that specific submatrices of it (defined by the so called structure) satisfy certain rank properties. Typically, a structured rank matrix $A$ has many submatrices with ranks much smaller than the matrix size and with sizes comparable to the size of the matrix. In that case we can call $A$ a *low-rank structured matrix*. In [33], Fiedler introduced the concept of *structure* and *structured rank* in the sense it will be used throughout this dissertation, and he studied different types of rank structures with regard to inversion and the $LU$-decomposition.

Here we will use a similar notation to that in [61]. In Definitions 2.1 and 2.2, which can be found in [61, Section 8.1], it is stated what is exactly meant by structured rank and low-rank structured matrices, respectively.

**Definition 2.1** (Structures, structured rank)**.** *Let $A$ be an $m \times n$ matrix. Denote with $\mathcal{M}$ the set of numbers $\{1, 2, \ldots, m\}$ and with $\mathcal{N}$ the set of numbers $\{1, 2, \ldots, n\}$, both with the usual order in $\mathbb{N}$. Let $\mathfrak{A}$ and $\mathfrak{B}$ be nonempty subsets of $\mathcal{M}$ and $\mathcal{N}$, respectively, also with the usual order in $\mathbb{N}$. Then, we denote by $A(\mathfrak{A}, \mathfrak{B})$ the subma-*

*trix of $A$ with row indices in $\mathfrak{A}$ and columns indices in $\mathfrak{B}$. A* structure $\Sigma$ *is defined as a nonempty subset of $\mathcal{M} \times \mathcal{N}$. Based on a given structure $\Sigma$, the* structured rank $r(\Sigma, A)$ *of $A$ is defined as*

$$\mathrm{r}(\Sigma, A) = \max\{\mathrm{rank}(A(\mathfrak{A}, \mathfrak{B}))\,|\mathfrak{A} \times \mathfrak{B} \subseteq \Sigma\}, \tag{2.1}$$

*where $\mathfrak{A} \times \mathfrak{B}$ denotes the set $\{(i,j)|i \in \mathfrak{A}, j \in \mathfrak{B}\}$.*

**Definition 2.2** (low-rank structured matrix)**.** *A matrix $A$ is called a* low-rank structured matrix *if there exists a structure $\Sigma$ of $A$ such that the structured rank* $\mathrm{r}(\Sigma, A)$ *is much smaller than the rank one would generically expect from the sizes of the submatrices that are determined by the structure $\Sigma$.*

In order to describe some examples of low-rank structured matrices, note that for any matrix $A$ of size $m \times n$ we can call its main diagonal $\{A(i,i),\ 1 \le i \le n\}$, as the 0th *diagonal* and therefore, for any natural number $p \ge 0$, we call the $p$-th *superdiagonal* of $A$, to the entries defined by $\{A(i, i+p),\ 1 \le i \le n-p\}$. In an analogous way, we call the $p$-th *subdiagonal* of $A$ to the entries defined by $\{A(i+p, i),\ 1 \le i \le n-p\}$.

Some examples of structures are presented in the following definition [61, p. 295].

**Definition 2.3** (Lower triangular structures)**.** *For a matrix $A$, and for $\mathcal{M} = \{1, 2, \ldots, m\}$ and $\mathcal{N} = \{1, 2, \ldots, n\}$ we define the following structures:*

- *The subset*
$$\Sigma_l = \{(i,j)|i \ge j, i \in \mathcal{M}, j \in \mathcal{N}\}$$
*is called* the lower triangular structure*, since the elements of this structure correspond to the indices from the lower triangular part of the matrix.*

- *The subset*
$$\Sigma_{wl} = \{(i,j)|i > j, i \in \mathcal{M}, j \in \mathcal{N}\}$$
*is called* the strictly lower triangular structure*.*

- *The subset*
$$\Sigma_l^{(p)} = \{(i,j)|i > j - p, i \in \mathcal{M}, j \in \mathcal{N}\}$$
*is called* the $p$-lower triangular structure*, and corresponds to the indices of all the entries of the matrix $A$ below the $p$-th diagonal. The 0th diagonal corresponds to the main diagonal, while the $p$-th diagonal refers to the $p$-th superdiagonal (for $p > 0$) and the $-p$-th diagonal refers to the $p$-th subdiagonal (for $p > 0$).*

Note that the $p$-th lower triangular structure for $p > 1$ contains not only the indices of the lower triangular part of the matrix but also the entries of the superdiagonals below the $p$-th superdiagonal. In an analogous way to Definition 2.3, we can define the structures $\Sigma_u$, $\Sigma_{wu}$ and $\Sigma_u^{(p)}$ for the upper triangular part of the matrix $A$.

As one would expect, it is obvious from Definition 2.3 that any tridiagonal matrix is a very simple example of a low-rank structured matrix since a matrix $A$ is tridiagonal if and only if:

$$\mathrm{r}\left(\Sigma_l^{(-1)}, A\right) = 0 \ \text{ and } \ \mathrm{r}\left(\Sigma_u^{(1)}, A\right) = 0.$$

Furthermore, according to Definition 2.3, all banded matrices are low-rank structured since they can be defined as follows.

**Definition 2.4.** *An $n \times n$ matrix $B$ is called a* $\{p, q\}$-banded matrix*, with $p \geq 0$ and $q \geq 0$, if the following two properties are satisfied:*

$$\mathrm{r}\left(\Sigma_l^{-p}, B\right) = 0 \text{ and } \mathrm{r}\left(\Sigma_u^{q}, B\right) = 0.$$

Obviously, the definition above means that a matrix $B$ is a $\{p, q\}$-banded matrix if and only if all of its entries below the $p$-th subdiagonal, and above the $q$-th superdiagonal, are equal to zero.

The examples above correspond to sparse matrices. On the other hand, a low-rank structured matrix may not be sparse. In fact, all the banded matrices are just particular sparse examples of a more general class of low-rank structured matrices named *semiseparable matrices*, which are generically dense.

The following definition can be found in [61, p. 300].

**Definition 2.5** ($\{n_L, n_U\}$-semiseparable matrix)**.** *A matrix $A \in \mathbb{R}^{n \times n}$ is called an* $\{n_L, n_U\}$-semiseparable matrix*, with $n_L \geq 0$ and $n_U \geq 0$, if the following two properties are satisfied:*

- *every submatrix of $A$ entirely located below the $n_L$th superdiagonal of $A$ has rank at most $n_L$, and there is at least one of these submatrices which has rank equal to $n_L$.*
- *every submatrix of $A$ entirely located above the $n_U$th subdiagonal of $A$ has rank at most $n_U$, and there is at least one of these submatrices which has rank equal to $n_U$.*

*This is obviously equivalent, to*

- $\max_i \text{ rank } A(i : n, 1 : i + n_L - 1) = n_L$, $\forall i \leq n - n_L + 1$, *and*
- $\max_i \text{ rank } A(1 : i + n_U - 1, i : n) = n_U$, $\forall i \leq n - n_U + 1$.

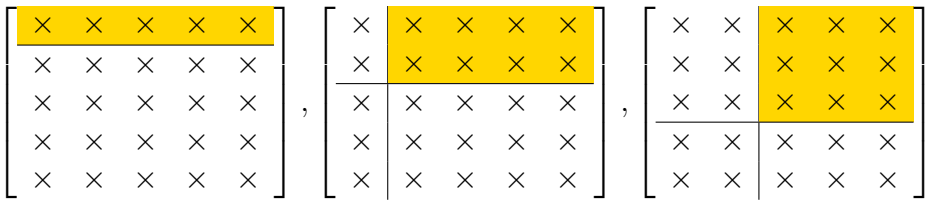
Note that, in the notation of Definition 2.3 and from Definition 2.5, a matrix $A$ is an $\{n_L, n_U\}$-semiseparable matrix if and only if $r\left(\Sigma_l^{(n_L)}, A\right) = n_L$ and $r\left(\Sigma_u^{(-n_U)}, A\right) = n_U$.

A $\{1,1\}$-semiseparable matrix is also often referred to as a $\{1\}$-*semiseparable* matrix or simply as a *semiseparable* matrix.

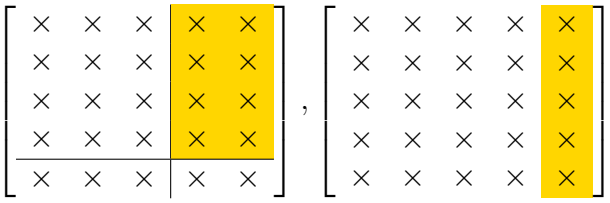
Graphically, if we denote by $\times$ any possible entry of the matrix (note that, in general, $\times$ has different numerical values for different entries), then for any matrix of size $5 \times 5$,

$$A = \begin{bmatrix} \times & \times & \times & \times & \times \\ \times & \times & \times & \times & \times \\ \times & \times & \times & \times & \times \\ \times & \times & \times & \times & \times \\ \times & \times & \times & \times & \times \end{bmatrix},$$

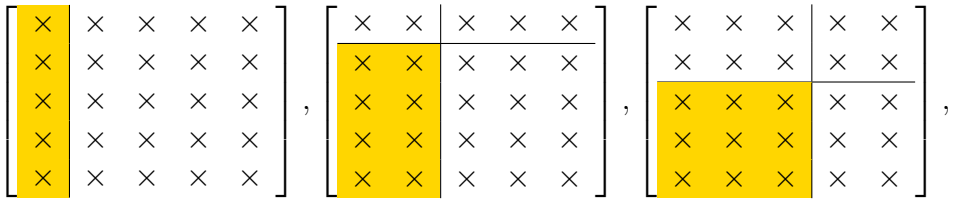
we have that $A$ is a $\{1,1\}$-semiseparable matrix if and only if all of the following marked submatrices corresponding to the upper triangular part of $A$ have rank at most 1,

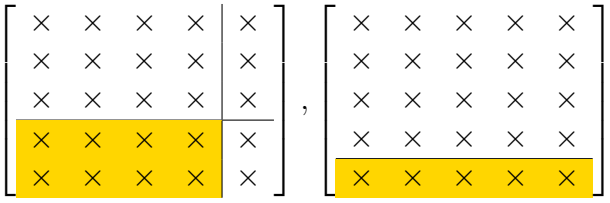


,



and all of the following marked submatrices corresponding to the lower triangular part of $A$ must also have rank at most 1:





Some other important classes of low rank structured matrices studied in [61, 62] are listed below in Definitions 2.6, 2.7, and 2.8. We will use some notation from Matlab, for any matrix $A \in \mathbb{R}^{n\times n}$, and for any natural number $0 \leq p \leq n-1$ we denote $\mathrm{tril}(A,p)$ to the lower triangular part of $A$ below and including the $p$-th subdiagonal, and $\mathrm{triu}(A,p)$ to the upper triangular part of $A$ above and including the $p$-th superdiagonal.

**Definition 2.6** (generator representable semiseparable matrix)**.** *A matrix $A \in \mathbb{R}^{n\times n}$ is called a generator representable semiseparable matrix, if the lower and upper tridiagonal parts of the matrix are coming from a rank 1 matrix, i.e., if the following two properties are satisfied:*

$$\mathrm{tril}(A,0) = \mathrm{tril}(\boldsymbol{u}\boldsymbol{v}^T) \;\; \textit{and} \;\; \mathrm{triu}(A,0) = \mathrm{tril}(\boldsymbol{p}\boldsymbol{q}^T),$$

*where $\boldsymbol{u},\boldsymbol{v},\boldsymbol{p},\boldsymbol{q} \in \mathbb{R}^{n\times 1}$, such that $\boldsymbol{u}_i\boldsymbol{v}_i = \boldsymbol{p}_i\boldsymbol{q}_i$, for all $i = 1,\ldots,n$.*

The generator definition has been used sometimes as a definition for semiseparable matrices [15, 31, 16, 64] (it is straigthforward from Definitions 2.6 and 2.5 that a generator representable semiseparable matrix is also a semiseparable matrix), but there are semiseparable matrices which are not generator representable (consider, for instance, any diagonal matrix with all its diagonal entries different from zero).

**Definition 2.7** (semiseparable plus diagonal matrix)**.** *A matrix $A \in \mathbb{R}^{n\times n}$ is called a semiseparable plus diagonal matrix if it can be expressed as the sum of a $\{1,1\}$-semiseparable and a diagonal matrix.*

These semiseparable plus diagonal matrices are related, for example, with the discretization of particular integral equations (see [47, 48] and [61, Section 3.3]).

**Definition 2.8** ($\{n_L,n_U\}$-generator representable semiseparable matrix)**.** *A matrix $A \in \mathbb{R}^{n\times n}$ is called an* $\{n_L,n_U\}$-generator representable semiseparable matrix*, with $n_L \geq 1$ and $n_U \geq 1$, if the following two properties are satisfied:*

$$\begin{aligned}\mathrm{tril}(A, n_L-1) &= \mathrm{tril}(UV^T, n_L-1),\\ \mathrm{triu}(A, -n_U+1) &= \mathrm{triu}(PQ^T, -n_U+1),\end{aligned}$$

*where $U,V \in \mathbb{R}^{n\times n_L}$ and $P,Q \in \mathbb{R}^{n\times n_U}$.*

Note that $\{n_L,n_U\}$-generator representable semiseparable matrices are a generalization of the matrices in Definition 2.6, and that both of them, as well as the semiseparable plus diagonal matrices in Definition 2.7, are particular cases of semiseparable matrices with different orders of semiseparability.

## 2.2 Quasiseparable matrices

According to Definition 2.5, it is obvious that the structure of a semiseparable matrix *crosses*, in general, its diagonal. On the other hand, if for $n \times n$ matrices we only consider strictly lower triangular and strictly upper triangular structures, then we have a more general class of low-rank structured matrices, since for any natural numbers $1 \leq p < n$ and $1 \leq q < n$, any submatrix contained in the strictly lower triangular part or in the upper triangular part of a matrix is also contained in the $p$-lower triangular part or in the $q$-upper triangular part of the matrix, respectively, and therefore, any constrain in the rank of the submatrices entirely located in the $p$-lower triangular part or in the $q$-upper triangular part of a matrix, respectively, is also a constraint in the rank of the submatrices in the strictly lower or upper triangular parts of the matrix, respectively. This more general class of low-rank structured matrices is named *quasiseparable* matrices, and we are specially interested on them in this thesis. The following definition can be found in [61, p. 301].

**Definition 2.9** ($\{n_L, n_U\}$-quasiseparable matrix)**.** *A matrix $A \in \mathbb{R}^{n\times n}$ is called an* $\{n_L, n_U\}$-quasiseparable matrix*, with $n_L \geq 0$ and $n_U \geq 0$, if the following two properties are satisfied:*

- *every submatrix of $A$ entirely located in the strictly lower triangular part of $A$ has rank at most $n_L$, and there is at least one of these submatrices which has rank equal to $n_L$, and*
- *every submatrix of $A$ entirely located in the strictly upper triangular part of $A$ has rank at most $n_U$, and there is at least one of these submatrices which has rank equal to $n_U$.*

*This is obviously equivalent to*

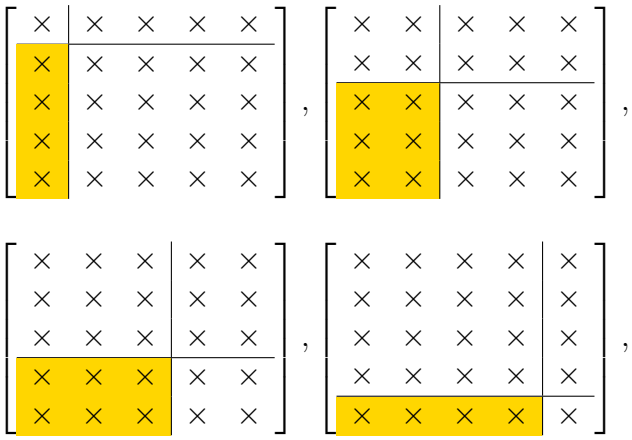
- $\max_i \text{ rank } A(i+1:n, 1:i) = n_L$*, and*
- $\max_i \text{ rank } A(1:i, i+1:n) = n_U$*.*

Note that, in the notation of Definition 2.3 and from Definition 2.9, a matrix $A$ is an $\{n_L, n_U\}$-quasiseparable matrix if and only if $\mathrm{r}(\Sigma_{wl}, A) = n_L$ and $\mathrm{r}(\Sigma_{wu}, A) = n_U$.

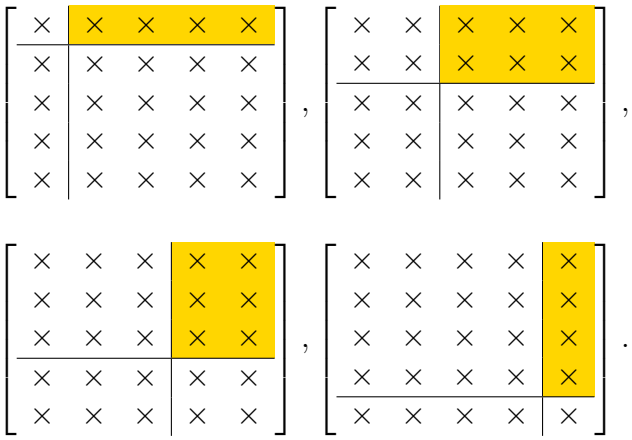
A $\{1,1\}$-quasiseparable matrix is often referred to as a $\{1\}$-*quasiseparable* matrix or simply as a *quasiseparable* matrix. The next figures illustrate graphically the preceding definition for any matrix $A$ of size $5 \times 5$,

$$A = \begin{bmatrix} \times & \times & \times & \times & \times \\ \times & \times & \times & \times & \times \\ \times & \times & \times & \times & \times \\ \times & \times & \times & \times & \times \\ \times & \times & \times & \times & \times \end{bmatrix}.$$

$A$ is an $\{n_L, n_U\}$-quasiseparable matrix, if and only if the following marked submatrices of $A$ have rank at most $n_L$,



and the following marked submatrices of $A$ have rank at most $n_U$,



We conclude this section by remarking that from Definitions 2.5 and 2.9, every $\{n_L, n_U\}$-semiseparable matrix is again an $\{n_L, n_U\}$-quasiseparable matrix (and also are all the particular cases of semiseparable matrices in Definitions 2.6, 2.7, and 2.8), but that the opposite is not true.

### 2.2.1 Inverses of quasiseparable matrices

A very useful result in the study of the relations between different types of structured rank matrices and their inverses was stated for the first time in [40] and it is known as the *Nullity Theorem*. We will present this theorem in the way it was formulated by Fiedler in [34].

**Definition 2.10.** *Given a matrix $A \in \mathbb{R}^{m\times n}$, we define the* nullity *of $A$, denoted* $\mathrm{n}(A)$*, as the dimension of the right null space of $A$.*

**Theorem 2.11.** *Let $A \in \mathbb{R}^{n\times n}$ be a nonsingular matrix, and denote by $B$ its inverse. Consider the following partition of $A$:*

$$A = \begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix},$$

*with $A_{11}$ of size $p \times q$, and the following partition of $B$:*

$$B = \begin{bmatrix} B_{11} & B_{12} \\ B_{21} & B_{22} \end{bmatrix},$$

*with $B_{11}$ of size $q \times p$. Then the nullities $\mathrm{n}(A_{11})$ and $\mathrm{n}(B_{22})$ are equal.*

*Proof.* Let us suppose first that $\mathrm{n}(B_{22}) \geq \mathrm{n}(A_{11})$. In this case, if $\mathrm{n}(B_{22}) = 0$ then $\mathrm{n}(A_{11}) = 0$ and the theorem is true. Therefore let us consider $\mathrm{n}(B_{22}) = c > 0$. Then we can construct a matrix $F \in \mathbb{R}^{(n-p)\times c}$ with its $c$ columns linearly independent and such that $B_{22}F = 0$. Note that from the identity $AB = I$ we can obtain the following equations:

$$A_{11}B_{12} + A_{12}B_{22} = 0. \tag{2.2}$$

$$A_{21}B_{12} + A_{22}B_{22} = I. \tag{2.3}$$

Note that if we multiply equations (2.2) and (2.3) on the right by $F$, we obtain:

$$A_{11}B_{12}F = 0, \tag{2.4}$$

$$A_{21}B_{12}F = F, \tag{2.5}$$

respectively. Hence, from (2.4) we conclude that $\mathrm{n}(A_{11}) \geq \mathrm{rank}(B_{12}F)$ and, from (2.5), we conclude that $\mathrm{rank}(B_{12}F) \geq \mathrm{rank}(F) = c$. Thus, we have obtained that

$$\mathrm{n}(A_{11}) \geq \mathrm{rank}(B_{12}F) \geq c = \mathrm{n}(B_{22}).$$

This last equation, together with the initial assumption $\mathrm{n}(B_{22}) \geq \mathrm{n}(A_{11})$, proves the theorem for this case.

On the other hand, if $\mathrm{n}(B_{22}) \leq \mathrm{n}(A_{11})$, we can consider the matrix

$$C = \begin{bmatrix} B_{22} & B_{21} \\ B_{12} & B_{11} \end{bmatrix},$$

and note that its inverse is given by

$$D = \begin{bmatrix} A_{22} & A_{21} \\ A_{12} & A_{11} \end{bmatrix}.$$

Then, since $\mathrm{n}(D_{22}) = \mathrm{n}(A_{11}) \geq \mathrm{n}(B_{22}) = \mathrm{n}(C_{11})$, we are in the previous case and we can conclude that $\mathrm{n}(A_{11}) = \mathrm{n}(B_{22})$, which completes the proof of the theorem. □

**Corollary 2.12.** *Let $A \in \mathbb{R}^{n\times n}$ and $B \in \mathbb{R}^{n\times n}$ be matrices such that $B = A^{-1}$, and consider the following partitions:*

$$A = \begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix}, \; B = \begin{bmatrix} B_{11} & B_{12} \\ B_{21} & B_{22} \end{bmatrix},$$

*with $A_{11}$ of size $p \times q$ and $B_{11}$ of size $q \times p$. Then, we have*

$$n(A_{12}) = \mathrm{n}(B_{12}) \text{ and } \mathrm{n}(A_{21}) = \mathrm{n}(B_{21}).$$

*Proof.* Consider the matrix $P$ given by the following partition

$$P = \begin{bmatrix} 0 & I_q \\ I_{n-q} & 0 \end{bmatrix},$$

where $I_q$ denotes the identity matrix of size $q \times q$, and note that the matrices

$$P^T B = \begin{bmatrix} B_{21} & B_{22} \\ B_{11} & B_{12} \end{bmatrix} \text{ and } AP = \begin{bmatrix} A_{12} & A_{11} \\ A_{22} & A_{21} \end{bmatrix},$$

are also each other's inverse. Then, according to Theorem 2.11, it follows that $\mathrm{n}(A_{12}) = \mathrm{n}(B_{12})$ and $\mathrm{n}(A_{21}) = \mathrm{n}(B_{21})$. □

An important consequence of the nullity theorem was formulated in [34] and it is stated in the following theorem. This is a useful result for proving rank properties of the inverses of structured rank matrices. In the rest of this section we use a standard notation for sets since for any two sets of natural numbers $\mathcal{A}$ and $\mathcal{B}$ we denote by $|\mathcal{A}|$ the cardinality of $\mathcal{A}$, i.e., $|\mathcal{A}|$ is the number of elements in $\mathcal{A}$, and $\mathcal{A} \setminus \mathcal{B}$ denotes the difference operation for sets, i.e., $\mathcal{A} \setminus \mathcal{B}$ is the set of elements of $\mathcal{A}$ that are not in $\mathcal{B}$.

**Theorem 2.13.** *Let $A \in \mathbb{R}^{n\times n}$ be an invertible matrix, and $\mathcal{A}, \mathcal{B}$ be subsets of $N = \{1, 2, ..., n\}$ such that their respective cardinalities $|\mathcal{A}|$ and $|\mathcal{B}|$ are both smaller than $n$. Then*

$$\mathrm{rank}\left(A^{-1}(\mathcal{A}, \mathcal{B})\right) = \mathrm{rank}\left(A(N \setminus \mathcal{B}, N \setminus \mathcal{A})\right) + |\mathcal{A}| + |\mathcal{B}| - n.$$

*Proof.* By permuting its rows and columns, the matrix $A$ can be transformed into a matrix $A'$ such that if we denote by $B'$ its inverse, they both can be partitioned as in the nullity theorem in a way such that the upper left submatrix $A'_{11}$ of $A'$ coincides with the submatrix $A(N \setminus \mathcal{B}, N \setminus \mathcal{A})$ of $A$ and, consequently, the lower right submatrix $B'_{22}$ of $B'$ coincides with the submatrix $B(\mathcal{A}, \mathcal{B})$ of the inverse $B$ of $A$. Then, using the nullity theorem, we have

$$\mathrm{n}(A'_{11}) = \mathrm{n}(B'_{22}). \tag{2.6}$$

On the other hand, since $A'_{11}$ is a matrix of dimension $N \setminus \mathcal{B} \times N \setminus \mathcal{A}$ and $B'_{22}$ has dimension $\mathcal{A} \times \mathcal{B}$, the following equalities hold:

$$\mathrm{n}(A'_{11}) = n - |\mathcal{A}| - \mathrm{rank}(A'_{11}), \tag{2.7}$$

$$\mathrm{n}(B'_{22}) = |\mathcal{B}| - \mathrm{rank}(B'_{22}). \tag{2.8}$$

From (2.6), (2.7), and (2.8), we end the proof of this theorem. □

**Corollary 2.14.** *Let $A \in \mathbb{R}^{n\times n}$ be an invertible matrix, and $\mathcal{A}$ be a subset of $N = \{1, 2, ..., n\}$ such that its cardinality $|\mathcal{A}|$ is smaller than $n$. Then*

$$\mathrm{rank}\left(A^{-1}(\mathcal{A}, N \setminus \mathcal{A})\right) = \mathrm{rank}\left(A(\mathcal{A}, N \setminus \mathcal{A})\right).$$

*Proof.* By choosing $N \setminus \mathcal{B} = \mathcal{A}$ in Theorem 2.13, the result follows. □

This last corollary states, in particular, that the ranks of all the submatrices entirely located below or above the diagonal of an invertible matrix $A$ will be the same as the rank of the corresponding submatrices, in the same position, of the inverse $A^{-1}$.

Theorem 2.15 states the relation existing between the $\{n_L, n_U\}$-quasiseparable matrices and their inverses.

**Theorem 2.15.** *The inverse of an invertible $\{n_L, n_U\}$-quasiseparable matrix is again an $\{n_L, n_U\}$-quasiseparable matrix.*

*Proof.* This is a direct consequence of Corollary 2.14. □

Theorem 2.15 establishes a key difference between quasiseparable matrices and other classes of rank structured matrices, since the set of quasiseparable matrices is closed under inversion. This property has important consequences in different applications, as for instance, in the fast solution of linear systems whose coefficient matrices are quasiseparable.

## 2.3 Representations

Structured matrices can often be represented by sets of parameters different than the sets of their entries. As one would expect, these representations are especially useful when they involve a much smaller number of parameters than the number of entries of the matrix. In some cases like banded matrices, it is straightforward to find such a representation, take, for instance, the set of all the entries of the matrix that are non identically zero and organize them in a way that their position in the matrix it is known, but, in general, representations are not always easy to find. In Definition 2.16 (see [61, p. 56]) we will state what is exactly meant by a representation of a class of matrices.

**Definition 2.16.** *Let* V *and* W *be vector spaces containing the sets* $\mathcal{V}$ *and* $\mathcal{W}$ *respectively, and such that* $\dim(\mathrm{V}) \leq \dim(\mathrm{W})$. *An element* $v \in \mathcal{V}$ *is said to be a* representation *of another element* $w \in \mathcal{W}$ *if there exists a map* $r$

$$r : \mathcal{V} \longrightarrow \mathcal{W},$$

*such that*

$$r(v) = w, \text{ and } r(\mathcal{V}) = \mathcal{W}.$$

*The map* $r$ *is called a representation map of the set* $\mathcal{W}$.

It is important to remark that the definition of a representation involves its existence not only for a single element $w$ but for the whole given subset $\mathcal{W}$. In practical situations, the knowledge of a map $s : \mathcal{W} \longrightarrow \mathcal{V}$ such that the map $r \circ s = r(s) : \mathcal{W} \longrightarrow \mathcal{W}$ is bijective and $r(s(w)) = w, \forall w \in \mathcal{W}$, is also needed, since this map will allow us to obtain the desired representation for any element $w \in \mathcal{W}$. We may also note from the previous definition that a representation of a given element $w \in \mathcal{W}$ (consider for instance a matrix in a given matrix class) may not be unique and, consequently, the choice of a representation for such class of matrices will heavily depend in criteria such as the number of parameters used by the representations and its stability with respect to the specific problem involving such matrices that we would like to solve, etc. An extensive description of different useful representation of low-rank structured matrices like semiseparable and quasiseparable matrices (and many other classes) can be found in the book by Vandebril, van Barel and Mastronardi [61, I.2, III.8.5].

In order to make Definition 2.16 clear, a representation for the very well known class of Vandermonde matrices is given in Example 2.17.

**Example 2.17.** *Let* $\mathcal{W} \subset \mathbb{R}^{m \times n}$ *be the set of all the* Vandermonde matrices *of size* $m \times n$. *Then, the map* $r : \mathbb{R}^m \longrightarrow \mathcal{W}$, *such that:*

$$r(\alpha_1, \alpha_2, \ldots, \alpha_m) = \begin{bmatrix} 1 & \alpha_1 & \alpha_1^2 & \cdots & \alpha_1^{n-1} \\ 1 & \alpha_2 & \alpha_2^2 & \cdots & \alpha_2^{n-1} \\ 1 & \alpha_3 & \alpha_3^2 & \cdots & \alpha_3^{n-1} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & \alpha_m & \alpha_m^2 & \cdots & \alpha_m^{n-1} \end{bmatrix}, \tag{2.9}$$

*clearly satisfies the conditions in Definition 2.16 (recall that any* $m \times n$ *Vandermonde matrix can be defined by the expression in the right-hand side of (2.9)). Therefore,* $r$ *can be considered as a representation map of the set* $\mathcal{W}$ *of Vandermonde matrices. In addition, the map* $s$ *described in the paragraph below Definition 2.16 can be easily defined as the inverse of* $r$, *i.e.,* $s = r^{-1}$.

### 2.3.1 The quasiseparable representation for $\{1,1\}$-quasiseparable matrices

Theorem 2.18, stated for $\{1,1\}$-quasiseparable matrices, is a particular case of a theorem proved in [24] for $\{n_L, n_U\}$-quasiseparable matrices and shows how any $\{1,1\}$-quasiseparable matrix of size $n \times n$ can be represented with $\mathcal{O}(n)$ parameters instead of its $n^2$ entries. The reader can find the theorem for general $\{n_L, n_U\}$-quasiseparable matrices in Section 2.3.3.

**Theorem 2.18.** *A matrix $A \in \mathbb{R}^{n\times n}$ is a $\{1,1\}$-quasiseparable matrix if and only if it can be parameterized in terms of the following set of $7n-8$ real parameters,*

$$\Omega_{QS} = (\{p_i\}_{i=2}^{n}, \{a_i\}_{i=2}^{n-1}, \{q_i\}_{i=1}^{n-1}, \{d_i\}_{i=1}^{n}, \{g_i\}_{i=1}^{n-1}, \{b_i\}_{i=2}^{n-1}, \{h_i\}_{i=2}^{n}),$$

*as follows:*

$$A = \begin{bmatrix} d_1 & g_1h_2 & g_1b_2h_3 & \cdots & g_1b_2\ldots b_{n-1}h_n \\ p_2q_1 & d_2 & g_2h_3 & \cdots & g_2b_3\ldots b_{n-1}h_n \\ p_3a_2q_1 & p_3q_2 & d_3 & \cdots & g_3b_4\ldots b_{n-1}h_n \\ p_4a_3a_2q_1 & p_4a_3q_2 & p_4q_3 & \cdots & g_4b_5\ldots b_{n-1}h_n \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ p_na_{n-1}a_{n-2}\ldots a_2q_1 & p_na_{n-1}\ldots a_3q_2 & p_na_{n-1}\ldots a_4q_3 & \cdots & d_n \end{bmatrix},$$

*or, in a more compact notation,*

$$A = \begin{bmatrix} d_1 & & & \\ & d_2 & & g_ib_{ij}^{\times}h_j \\ p_ia_{ij}^{\times}q_j & & \ddots & \\ & & & d_n \end{bmatrix},$$

*where $a_{ij}^{\times} = a_{i-1}a_{i-2}\cdots a_{j+1}$, for $i-1 \geq j+1$, $b_{ij}^{\times} = b_{i+1}b_{i+2}\cdots b_{j-1}$, for $i+1 \leq j-1$, $a_{j+1,j}^{\times} = 1$, and $b_{j,j+1}^{\times} = 1$ for $j = 1,\ldots,n-1$.*

Let us denote by $\mathrm{Q}^n \subset \mathbb{R}^{n\times n}$ the set of all $\{1,1\}$-quasiseparable matrices of size $n \times n$. From Definition 2.16 and the previous theorem, the set of parameters $\Omega_{QS}$ is a representation of the matrix $A \in \mathrm{Q}^n$ and therefore we call $\Omega_{QS}$ a *quasiseparable representation* of $A$. Note that this representation is not unique as we can see in the following example.

**Example 2.19.** *Let $A$ be a $\{1,1\}$-quasiseparable matrix of size $5 \times 5$ and consider a quasiseparable representation of $A$:*

$$\Omega_{QS} = (\{p_i\}_{i=2}^{5}, \{a_i\}_{i=2}^{4}, \{q_i\}_{i=1}^{4}, \{d_i\}_{i=1}^{5}, \{g_i\}_{i=1}^{4}, \{b_i\}_{i=2}^{4}, \{h_i\}_{i=2}^{5}).$$

*Then,*

$$A = \begin{bmatrix} d_1 & g_1h_2 & g_1b_2h_3 & g_1b_2b_3h_4 & g_1b_2b_3b_4h_5 \\ p_2q_1 & d_2 & g_2h_3 & g_2b_3h_4 & g_2b_3b_4h_5 \\ p_3a_2q_1 & p_3q_2 & d_3 & g_3h_4 & g_3b_4h_5 \\ p_4a_3a_2q_1 & p_4a_3q_2 & p_4q_3 & d_4 & g_4h_5 \\ p_5a_4a_3a_2q_1 & p_5a_4a_3q_2 & p_5a_4q_3 & p_5q_4 & d_5 \end{bmatrix},$$

*and for every real number $\alpha \neq 0, 1$, we also have*

$$A = \begin{bmatrix} d_1 & g_1h_2 & g_1b_2h_3 & g_1b_2b_3h_4 & g_1b_2b_3b_4h_5 \\ (\alpha p_2)\frac{q_1}{\alpha} & d_2 & g_2h_3 & g_2b_3h_4 & g_2b_3b_4h_5 \\ (\alpha p_3)a_2\frac{q_1}{\alpha} & (\alpha p_3)\frac{q_2}{\alpha} & d_3 & g_3h_4 & g_3b_4h_5 \\ (\alpha p_4)a_3a_2\frac{q_1}{\alpha} & (\alpha p_4)a_3\frac{q_2}{\alpha} & (\alpha p_4)\frac{q_3}{\alpha} & d_4 & g_4h_5 \\ (\alpha p_5)a_4a_3a_2\frac{q_1}{\alpha} & (\alpha p_5)a_4a_3\frac{q_2}{\alpha} & (\alpha p_5)a_4\frac{q_3}{\alpha} & (\alpha p_5)\frac{q_4}{\alpha} & d_5 \end{bmatrix},$$

*and we obtain a different quasiseparable representation of $A$:*

$$\Omega'_{QS} = (\{\alpha p_i\}_{i=2}^5, \{a_i\}_{i=2}^4, \{q_i/\alpha\}_{i=1}^4, \{d_i\}_{i=1}^5, \{g_i\}_{i=1}^4, \{b_i\}_{i=2}^4, \{h_i\}_{i=2}^5).$$

*There are many other ways in which the quasiseparable representation may be not unique.*

It is worth to mention that the class of quasiseparable matrices is often defined in terms of its representations. In [61, Ch. 2, Sec. 8.5], the reader can find an extensive description of different useful representations for quasiseparable matrices and for some other low-rank structured matrices.

**Remark 2.20.** *There are many important subsets of $\{1,1\}$-quasiseparable matrices arising in applications as, for instance, semiseparable matrices, generator representable semiseparable matrices, semiseparable plus diagonal matrices, and their corresponding symmetric versions [61, Ch. 1]. Although these particular subsets of matrices can be represented via the quasiseparable representation introduced in Theorem 2.18, they also admit other "more compressed" representations, i.e., in terms of less parameters, which are special instances of the quasiseparable representation. Such compressed representations can be found in [61, Chs. 1 and 2] and are the ones to be used in practice when working with these particular $\{1,1\}$-quasiseparable matrices. The formalism presented later in Chapters 4 and 5 of this thesis can be directly applied to develop condition numbers with respect to these compressed representations. These condition numbers would reflect faithfully the particular structures of the subsets of matrices mentioned above and, therefore, would be smaller than the condition numbers developed in Chapters 4 and 5, since they restrict the possible perturbations in order to preserve the additional structures. For the sake of brevity, we do not develop such condition numbers in this thesis.*

### 2.3.2 The Givens-vector representation for $\{1,1\}$-quasiseparable matrices

Another important representation for $\{1,1\}$-quasiseparable matrices is the Givens-vector representation introduced in [59]. This representation was introduced to improve the numerical stability of fast matrix computations involving quasiseparable matrices with respect to other representations, but a rigorous proof that this is indeed the case has never been given. The results in this dissertation are a first contribution to the solution of this problem (see Sections 4.3 and 5.5 for results on the sensitivities of the solution of linear systems whose coefficient matrices are $\{1,1\}$-quasiseparable and of eigenvalues of $\{1,1\}$-quasiseparable matrices, respectively). The next theorem shows that this representation is able of representing the complete class of $\{1,1\}$-quasiseparable matrices (see [61, Sections 2.4 and 2.8]).

**Theorem 2.21.** *A matrix $A \in \mathbb{R}^{n\times n}$ is a $\{1,1\}$-quasiseparable matrix if and only if it can be parameterized in terms of the following set of parameters,*

- $\{c_i, s_i\}_{i=2}^{n-1}$, *where $(c_i, s_i)$ is a pair of cosine-sine with $c_i^2 + s_i^2 = 1$ for every $i \in \{2, 3, \cdots, n-1\}$,*
- $\{v_i\}_{i=1}^{n-1}, \{d_i\}_{i=1}^{n}, \{e_i\}_{i=1}^{n-1}$ *all of them independent real parameters,*
- $\{r_i, t_i\}_{i=2}^{n-1}$, *where $(r_i, t_i)$ is a pair of cosine-sine with $r_i^2 + t_i^2 = 1$ for every $i \in \{2, 3, \cdots, n-1\}$,*

*as follows:*

$$A = \begin{bmatrix} d_1 & e_1r_2 & e_1t_2r_3 & \cdots & e_1t_2\dots t_{n\ 2}r_{n\ 1} & e_1t_2\dots t_{n\ 1} \\ c_2v_1 & d_2 & e_2r_3 & \cdots & e_2t_3\dots t_{n\ 2}r_{n\ 1} & e_2t_3\dots t_{n\ 1} \\ c_3s_{2\ 1}v_1 & c_3v_2 & d_3 & \cdots & e_3t_4\dots t_{n\ 2}r_{n\ 1} & e_3t_4\dots t_{n\ 1} \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ c_{n\ 1}s_{n\ 2}\dots s_{2\ 1}v_1 & c_{n\ 1}s_{n\ 2}\dots s_{3\ 1}v_2 & c_{n\ 1}s_{n\ 2}\dots s_{4\ 1}v_3 & \cdots & d_{n\ 1} & e_{n\ 1} \\ s_{n\ 1}s_{n\ 2}\dots s_{2\ 1}v_1 & s_{n\ 1}s_{n\ 2}\dots s_{3\ 1}v_2 & s_{n\ 1}s_{n\ 2}\dots s_{4\ 1}v_3 & \cdots & v_{n\ 1} & d_n \end{bmatrix}.$$

*This representation is denoted by $\Omega_{QS}^{GV}$, i.e.,*

$$\Omega_{QS}^{GV} := \left(\{c_i, s_i\}_{i=2}^{n-1}, \{v_i\}_{i=1}^{n-1}, \{d_i\}_{i=1}^{n}, \{e_i\}_{i=1}^{n-1}, \{r_i, t_i\}_{i=2}^{n-1}\right).$$

From Theorems 2.18 and 2.21, it is obvious that the Givens-vector representation is a particular case of the quasiseparable representation for $\{1,1\}$-quasiseparable matrices by considering the following relations between the parameters in Theorems 2.18 and 2.21, respectively: $\{p_i, a_i\}_{i=2}^{n-1} = \{c_i, s_i\}_{i=2}^{n-1}$, $\{q_i\}_{i=1}^{n-1} = \{v_i\}_{i=1}^{n-1}$, $\{d_i\}_{i=1}^{n} = \{d_i\}_{i=1}^{n}$, $\{g_i\}_{i=1}^{n-1} = \{e_i\}_{i=1}^{n-1}$, $\{b_i, h_i\}_{i=2}^{n-1} = \{t_i, r_i\}_{i=2}^{n-1}$, and $p_n = h_n = 1$. This fact can be observed better by comparing the expression in Example 2.19 and the expression in the following $5 \times 5$ example.

**Example 2.22.** *Let $A \in \mathbb{R}^{5\times 5}$ be a $\{1,1\}$-quasiseparable matrix, and let*

$$\Omega_{QS}^{GV} := \left(\{c_i, s_i\}_{i=2}^{n-1}, \{v_i\}_{i=1}^{n-1}, \{d_i\}_{i=1}^{n}, \{e_i\}_{i=1}^{n-1}, \{r_i, t_i\}_{i=2}^{n-1}\right)$$

*be a Givens-vector representation of $A$. Then,*

$$A = \begin{bmatrix} d_1 & e_1r_2 & e_1t_2r_3 & e_1t_2t_3r_4 & e_1t_2t_3t_4 \\ c_2v_1 & d_2 & e_2r_3 & e_2t_3r_4 & e_2t_3t_4 \\ c_3s_2v_1 & c_3v_2 & d_3 & e_3r_4 & e_3t_4 \\ c_4s_3s_2v_1 & c_4s_3v_2 & c_4v_3 & d_4 & e_4 \\ s_4s_3s_2v_1 & s_4s_3v_2 & s_4v_3 & v_4 & d_5 \end{bmatrix}.$$

Note that the Givens-vector representation can be made unique if $c_i$ and $r_i$ are taken to be nonnegative numbers (if $c_i = 0$, take $s_i = 1$ and if $r_i = 0$, take $t_i = 1$) [61, p.76].

### 2.3.3 The quasiseparable representation for $\{n_L, n_U\}$-quasiseparable matrices

The representation presented in Section 2.3.1 has a nice extension towards the more general class of the $\{n_L, n_U\}$-quasiseparable matrices, as stated in the next theorem from [24].

**Theorem 2.23.** *A matrix $A$ is an $\{n_L, n_U\}$-quasiseparable matrix if and only if it can be parameterized in terms of the set of parameters defined as the entries of the following matrices:*

$$\Omega_{QS} := (\{p_i\}_{i=2}^{n}, \{a_i\}_{i=2}^{n-1}, \{q_j\}_{j=1}^{n-1}, \{d_i\}_{i=1}^{n}, \{g_i\}_{i=1}^{n-1}, \{b_i\}_{i=2}^{n-1}, \{h_j\}_{j=2}^{n}),$$

*where:*

$\{d_i\}_{i=1}^{n} \subset \mathbb{R}^{1\times 1}$,
$\{p_i\}_{i=2}^{n} \subset \mathbb{R}^{1\times n_L}, \{a_i\}_{i=2}^{n-1} \subset \mathbb{R}^{n_L\times n_L}, \{q_j\}_{j=1}^{n-1} \subset \mathbb{R}^{n_L\times 1}$,
$\{g_i\}_{i=1}^{n-1} \subset \mathbb{R}^{1\times n_U}, \{b_i\}_{i=2}^{n-1} \subset \mathbb{R}^{n_U\times n_U}, \{h_j\}_{j=2}^{n} \subset \mathbb{R}^{n_U\times 1}$,

*as follows:*

$$A = \begin{bmatrix} d_1 & g_1h_2 & g_1b_2h_3 & \cdots & g_1b_2\ldots b_{n-1}h_n \\ p_2q_1 & d_2 & g_2h_3 & \cdots & g_2b_3\ldots b_{n-1}h_n \\ p_3a_2q_1 & p_3q_2 & d_3 & \cdots & g_3b_4\ldots b_{n-1}h_n \\ p_4a_3a_2q_1 & p_4a_3q_2 & p_4q_3 & \cdots & g_4b_5\ldots b_{n-1}h_n \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ p_na_{n-1}a_{n-2}\ldots a_2q_1 & p_na_{n-1}\ldots a_3q_2 & p_na_{n-1}\ldots a_4q_3 & \cdots & d_n \end{bmatrix}. \tag{2.10}$$

*In a more compact notation, we can write:*

$$A = \begin{bmatrix} d_1 & & & \\ & d_2 & & g_i b_{ij}^{\times} h_j \\ & & \ddots & \\ p_i a_{ij}^{\times} q_j & & & d_n \end{bmatrix},$$

*where* $a_{ij}^{\times} = a_{i-1}a_{i-2}\cdots a_{j+1}$, *for* $i-1 \geq j+1$, $b_{ij}^{\times} = b_{i+1}b_{i+2}\cdots b_{j-1}$, *for* $i+1 \leq j-1$, *and* $a_{j+1,j}^{\times} = 1$, $b_{j,j+1}^{\times} = 1$ *for* $j = 1, 2, \ldots, n-1$.

# Chapter 3

# Principles of condition numbers

This chapter is devoted to the definitions of the condition numbers that will be used troughout the rest of this thesis and to establish some of their main properties. In the first part of this chapter (Section 3.1), we study condition numbers for linear systems, while in the second part (Section 3.2) we study condition numbers for eigenvalues. The results in Sections 3.1.2 and 3.2.2 are original contributions of the author and will be respectively used in Chapters 4 and 5. More precisely, the most important results, in the context of this thesis, of this chapter are given in Theorems 3.10 and 3.25, where the expressions for the respective condition numbers for the solution of linear systems and for eigenvalues are deduced with respect to relative perturbations on the parameters generating the matrices involved in those problems.

## 3.1 Basics on condition numbers for linear systems

This section contains two subsections in order to separate the already known results from the new ones. Section 3.1.1 includes the standard definitions and results about unstructured condition numbers for the solution of linear systems, while in Section 3.1.2 we present some new definitions and results for structured condition numbers for the solution of linear systems whose coefficient matrices are differentiable functions of a set of parameters.

### 3.1.1 Standard results

We start this section by presenting some well-known results about condition numbers for the solution of a linear system of equations. Note first that any perturbation of a matrix $A \in \mathbb{R}^{n\times n}$ can be expressed as a sum $A + \delta A$, where $\delta A \in \mathbb{R}^{n\times n}$ is called the *perturbation matrix*, and let us define subordinate norms as in [41, Sec. 6.2].

**Definition 3.1.** *For any vector norm $\|\cdot\|$ on $\mathbb{C}^n$, the corresponding subordinate*

*matrix norm on* $\mathbb{C}^{m\times n}$ *is defined by*

$$\|A\| = \max_{x\neq 0} \frac{\|A\boldsymbol{x}\|}{\|\boldsymbol{x}\|},$$

*or, equivalently,*

$$\|A\| = \max_{\|x\|=1} \|A\boldsymbol{x}\|,$$

*where* $\boldsymbol{x} \in \mathbb{C}^n$ *and* $A \in \mathbb{C}^{m\times n}$.

Associated with a *normwise backward error* we have the condition number in Definition 3.2 [41, Sec. 7.1], valid for any vector norm and the corresponding subordinate matrix norm.

**Definition 3.2.** *Let* $A\boldsymbol{x} = \boldsymbol{b}$, *where* $A \in \mathbb{R}^{n\times n}$ *is nonsingular, and* $0 \neq \boldsymbol{x} \in \mathbb{R}^n$. *Then, for* $E \in \mathbb{R}^{n\times n}$ *and* $\boldsymbol{f} \in \mathbb{R}^n$, *we define*

$$\kappa_{E,\boldsymbol{f}}(A,\boldsymbol{x}) := \limsup_{\eta\to 0}\left\{ \frac{\|\delta\boldsymbol{x}\|}{\eta\|\boldsymbol{x}\|} : (A+\delta A)(\boldsymbol{x}+\delta\boldsymbol{x}) = \boldsymbol{b}+\delta\boldsymbol{b},\ \|\delta A\| \leq \eta\|E\|, \right.$$
$$\left. \|\delta\boldsymbol{b}\| \leq \eta\|\boldsymbol{f}\| \right\}.$$

Observe that $\kappa_{E,\boldsymbol{f}}(A,\boldsymbol{x})$ is a normwise relative condition number, i.e., it measures the relative sensitivity of the solution $\boldsymbol{x}$ of the linear system $A\boldsymbol{x} = \boldsymbol{b}$ with respect to relative normwise perturbations of the matrix and the right-hand side (note that, in this case, the perturbations are measured against the tolerances $E$ and $\boldsymbol{f}$). This condition number has the expression presented in the following theorem proved in [41, Sec. 7.1].

**Theorem 3.3.** *Under the same hypotheses of Definition 3.2,*

$$\kappa_{E,\boldsymbol{f}}(A,\boldsymbol{x}) = \frac{\|A^{-1}\|\|\boldsymbol{f}\|}{\|\boldsymbol{x}\|} + \|A^{-1}\|\|E\|.$$

Recall that the usual matrix condition number is given by $\kappa(A) := \|A\|\|A^{-1}\|$ and note that if we take $E = A$ and $\boldsymbol{f} = \boldsymbol{b}$, then we have $\kappa(A) \leq \kappa_{E,\boldsymbol{f}}(A,\boldsymbol{x}) \leq 2\kappa(A)$, and therefore they are numerically equivalent. On the other hand, it is well-known that considering normwise perturbations of the matrix $A$ and the vector $\boldsymbol{b}$ may lead to pessimistic bounds on the forward errors, since there are matrices and vectors for which we may have a small relative normwise perturbation that produces some large relative perturbations over their small entries, and that may affect the zero pattern of the matrix or the vector (see, for instance, the numerical example in [41, pp. 121,124]). Therefore, it makes sense to consider componentwise perturbations and the corresponding componentwise condition number. We denote by $|A|$ the matrix

whose entries are the absolute values of the entries of $A$ (i.e., $|A|_{ij} := |A_{ij}|$) and we adopt a similar notation for vectors. In addition, inequalities $|A| \leq |B|$ mean $|A_{ij}| \leq |B_{ij}|$ for all $i, j$. Definition 3.4 and Theorem 3.5 can both be found in [41, Sec. 7.2], together with a brief discussion about how to choose the tolerances $E$ and $\boldsymbol{f}$.

**Definition 3.4.** *Let $A\boldsymbol{x} = \boldsymbol{b}$, where $A \in \mathbb{R}^{n\times n}$ is nonsingular, and $0 \neq \boldsymbol{x} \in \mathbb{R}^n$. Then, for $0 \leq E \in \mathbb{R}^{n\times n}$ and $0 \leq \boldsymbol{f} \in \mathbb{R}^n$, we define the* relative componentwise condition number *as*

$$\operatorname{cond}_{E,\boldsymbol{f}}(A, \boldsymbol{x}) := \lim_{\eta\to 0} \sup \left\{ \frac{\|\delta\boldsymbol{x}\|_\infty}{\eta\|\boldsymbol{x}\|_\infty} : (A + \delta A)(\boldsymbol{x} + \delta\boldsymbol{x}) = \boldsymbol{b} + \delta\boldsymbol{b},\ |\delta A| \leq \eta E, \right.$$
$$\left. |\delta\boldsymbol{b}| \leq \eta\boldsymbol{f} \right\}.$$

**Theorem 3.5.** *Under the same hypotheses of Definition 3.4,*

$$\operatorname{cond}_{E,\boldsymbol{f}}(A, \boldsymbol{x}) = \frac{\||A^{-1}|E|\boldsymbol{x}| + |A^{-1}|\boldsymbol{f}\|_\infty}{\|\boldsymbol{x}\|_\infty}.$$

A proof of this theorem is provided in [41, Sec. 7.2], but it can be seen as a consequence of the more general Theorem 3.10 that we introduce in Section 3.1.2, and, so, we will present a proof of Theorem 3.5 at the end of that section.

From the expression in Theorem 3.5, if we consider $E = |A|$ and $f = |\boldsymbol{b}|$, then it is straightforward to prove that the condition number $\operatorname{cond}_{|A|,|\boldsymbol{b}|}(A, \boldsymbol{x})$ is invariant under row scaling. This useful property is stated in the following proposition.

**Proposition 3.6.** *Let $A\boldsymbol{x} = \boldsymbol{b}$, where $A \in \mathbb{R}^{n\times n}$ is nonsingular and $0 \neq \boldsymbol{x} \in \mathbb{R}^n$, and let $K \in \mathbb{R}^{n\times n}$ be an invertible diagonal matrix. Then, for $KA\boldsymbol{x} = K\boldsymbol{b}$, we have*

$$\operatorname{cond}_{|A|,|\boldsymbol{b}|}(A, \boldsymbol{x}) = \operatorname{cond}_{|KA|,|K\boldsymbol{b}|}(KA, \boldsymbol{x}).$$

### 3.1.2 Condition numbers for parameterized linear systems

Since many interesting classes of matrices can be represented by sets of parameters different from their entries (see Theorem 2.18, for example), we generalize the definitions in Section 3.1.1 to these representations and, following the ideas in [32, 42, 43], we will focus on componentwise relative condition numbers for representations.

**Definition 3.7.** *Let $A\boldsymbol{x} = \boldsymbol{b}$, where $A \in \mathbb{R}^{n\times n}$ is a nonsingular matrix whose entries are differentiable functions of a vector of parameters $\Omega = (\omega_1, \omega_2, \ldots, \omega_m)^T \in \mathbb{R}^m$, this is denoted by $A(\Omega)$, and $0 \neq \boldsymbol{x} \in \mathbb{R}^n$. Let $0 \leq \boldsymbol{f} \in \mathbb{R}^n$ and $E = (e_1, e_2, \ldots, e_m)^T \in \mathbb{R}^m$ with nonnegative entries. Then, we define*

$$\operatorname{cond}_{E,\boldsymbol{f}}(A(\Omega), \boldsymbol{x}) := \lim_{\eta\to 0} \sup \left\{ \frac{\|\delta\boldsymbol{x}\|_\infty}{\eta\|\boldsymbol{x}\|_\infty} : (A(\Omega + \delta\Omega))(\boldsymbol{x} + \delta\boldsymbol{x}) = \boldsymbol{b} + \delta\boldsymbol{b}, \right.$$

$$|\delta\Omega| \leq \eta E,\ |\delta \boldsymbol{b}| \leq \eta \boldsymbol{f} \Big\}.$$

The main goal of this section is to find an explicit expression for the componentwise relative condition number with respect to a general representation introduced in Definition 3.7. For such a purpose we will use differential calculus and we will need Lemma 3.8. Recall $\boldsymbol{e}_i$ denotes the $i$-th vector of the canonical basis of $\mathbb{R}^n$.

**Lemma 3.8.** *Let $A\boldsymbol{x} = \boldsymbol{b}$, where $A \in \mathbb{R}^{n\times n}$ is an invertible matrix whose entries are differentiable functions of a vector of real parameters $\Omega = (\omega_1, \omega_2, \ldots, \omega_m)^T$, and $0 \neq \boldsymbol{x} \in \mathbb{R}^n$. Then, the following equalities hold:*

(a) $\dfrac{\partial A^{-1}}{\partial \omega_k} = -A^{-1}\dfrac{\partial A}{\partial \omega_k}A^{-1}$, *for* $k \in \{1, 2, \ldots, m\}$,

(b) $\dfrac{\partial \boldsymbol{x}}{\partial b_i} = A^{-1}\boldsymbol{e}_i$, *for* $i \in \{1, 2, \ldots, n\}$,

(c) $\dfrac{\partial \boldsymbol{x}}{\partial \omega_k} = -A^{-1}\dfrac{\partial A}{\partial \omega_k}\boldsymbol{x}$, *for* $k \in \{1, 2, \ldots, m\}$.

*Proof.* (a) Derivating in both sides of $AA^{-1} = I_n$, we get

$$\frac{\partial A}{\partial \omega_k}A^{-1} + A\frac{\partial A^{-1}}{\partial \omega_k} = 0.$$

(b) It follows trivially from derivating $\boldsymbol{x} = A^{-1}\boldsymbol{b}$.

(c) From derivating $\boldsymbol{x} = A^{-1}\boldsymbol{b}$ and using (a), we obtain

$$\frac{\partial \boldsymbol{x}}{\partial \omega_k} = \frac{\partial A^{-1}}{\partial \omega_k}\boldsymbol{b} = -A^{-1}\frac{\partial A}{\partial \omega_k}A^{-1}\boldsymbol{b} = -A^{-1}\frac{\partial A}{\partial \omega_k}\boldsymbol{x}.$$

□

**Remark 3.9.** *In Lemma 3.8, we have used that the entries of $A^{-1}$ are also differentiable functions of $(\omega_1, \ldots, \omega_m)$. This follows from the facts that (1) each entry of $A^{-1}$ is a quotient of a cofactor of $A$ divided by $det(A)$ and that (2) products, sums, and quotients of differentiable functions are differentiable whenever the denominators are not zero.*

In Theorem 3.10, we provide the desired explicit expression of the componentwise relative condition number introduced in Definition 3.7.

**Theorem 3.10.** *Let $A\boldsymbol{x} = \boldsymbol{b}$, where $A \in \mathbb{R}^{n\times n}$ is an invertible matrix whose entries are differentiable functions of a vector of real parameters $\Omega = (\omega_1, \omega_2, \ldots, \omega_m)^T$ and $0 \neq \boldsymbol{x} \in \mathbb{R}^n$. Let $0 \leq \boldsymbol{f} \in \mathbb{R}^n$ and $E = (e_1, e_2, \ldots, e_m)^T \in \mathbb{R}^m$ with nonnegative entries. Then,*

$$\mathrm{cond}_{E,\boldsymbol{f}}(A(\Omega), \boldsymbol{x}) = \frac{\left\| |A^{-1}|\,\boldsymbol{f} + \sum_{k=1}^{m} \left| A^{-1}\dfrac{\partial A}{\partial \omega_k}\boldsymbol{x} \right| e_k \right\|_\infty}{\|\boldsymbol{x}\|_\infty}.$$

*Proof.* Since the entries of the matrix $A$ are differentiable functions of the parameters in $\Omega$ and, from $\boldsymbol{x} = A^{-1}\boldsymbol{b}$, it is clear that $\boldsymbol{x}$ is a function of $\Omega$ and $\boldsymbol{b}$, we can use differential calculus to obtain the following result:

$$\delta\boldsymbol{x} = \sum_{i=1}^{n} \frac{\partial \boldsymbol{x}}{\partial b_i}\delta b_i + \sum_{k=1}^{m} \frac{\partial \boldsymbol{x}}{\partial \omega_k}\delta\omega_k + \mathcal{O}(\|(\delta\Omega, \delta\boldsymbol{b})\|^2),$$

where $\|(\delta\Omega, \boldsymbol{b})\| := \max\{\|\delta\Omega\|_\infty, \|\delta\boldsymbol{b}\|_\infty\}$. Using (b) and (c) from Lemma 3.8 in the previous equation, we obtain:

$$\delta\boldsymbol{x} = \sum_{i=1}^{n} \left(A^{-1}\boldsymbol{e}_i\right)\delta b_i + \sum_{k=1}^{m} \left(-A^{-1}\frac{\partial A}{\partial \omega_k}\boldsymbol{x}\right)\delta\omega_k + \mathcal{O}(\|(\delta\Omega, \delta\boldsymbol{b})\|^2). \tag{3.1}$$

From (3.1), using standard properties of the $\infty$-norm and the inequalities $|\delta\boldsymbol{b}| \leq \eta\boldsymbol{f}$ and $|\delta\Omega| \leq \eta E$, we get

$$\begin{aligned}\|\delta\boldsymbol{x}\|_\infty &\leq \eta \left\| \sum_{i=1}^{n} \left|A^{-1}\boldsymbol{e}_i\right| f_i + \sum_{k=1}^{m} \left| A^{-1}\frac{\partial A}{\partial \omega_k}\boldsymbol{x} \right| e_k \right\|_\infty + \mathcal{O}(\|(\delta\Omega, \delta\boldsymbol{b})\|^2) \\ &= \eta \left\| \left|A^{-1}\right|\boldsymbol{f} + \sum_{k=1}^{m} \left| A^{-1}\frac{\partial A}{\partial \omega_k}\boldsymbol{x} \right| e_k \right\|_\infty + \mathcal{O}(\|(\delta\Omega, \delta\boldsymbol{b})\|^2). \end{aligned} \tag{3.2}$$

Then, if $\eta$ tends to zero, from (3.2) and from Definition 3.7, it is straightforward to get

$$\mathrm{cond}_{E,\boldsymbol{f}}(A(\Omega), \boldsymbol{x}) \leq \frac{\left\| \left|A^{-1}\right|\boldsymbol{f} + \sum_{k=1}^{m} \left| A^{-1}\dfrac{\partial A}{\partial \omega_k}\boldsymbol{x} \right| e_k \right\|_\infty}{\|\boldsymbol{x}\|_\infty}. \tag{3.3}$$

On the other hand, if we consider the perturbations:

$$\delta\boldsymbol{b} = \eta D\boldsymbol{f},$$

where $D$ is a diagonal matrix such that $D(j,j) = \mathrm{sign}\,(A^{-1}(l,j))$, for $j = 1, 2, \ldots, n$, and

$$\delta\omega_k = -\eta \left[ \mathrm{sign} \left( A^{-1}\frac{\partial A}{\partial \omega_k}\boldsymbol{x} \right)_l \right] e_k, \quad \text{for } k = 1, \ldots, m,$$

where $l$ is such that

$$\left\| \left|A^{-1}\right| \boldsymbol{f} + \sum_{k=1}^{m} \left| A^{-1} \frac{\partial A}{\partial \omega_k} \boldsymbol{x} \right| e_k \right\|_\infty = \left( \left|A^{-1}\right| \boldsymbol{f} + \sum_{k=1}^{m} \left| A^{-1} \frac{\partial A}{\partial \omega_k} \boldsymbol{x} \right| e_k \right)_l ,$$

we can obtain, from (3.1) and from Definition 3.7, the desired equality in (3.3). □

As a consequence of Theorem 3.10 we can deduce the very well-known expression in Theorem 3.5 for the condition number $\mathrm{cond}_{E,\boldsymbol{f}}(A, \boldsymbol{x})$ by considering $\Omega$ as *a vectorization* of the entries of $A$. Therefore, we conclude this section by providing its proof.

*Proof.* (of Theorem 3.5 ) Note first that, in this case, we can rewrite the expression in Theorem 3.10 as:

$$\mathrm{cond}_{E,\boldsymbol{f}}(A, \boldsymbol{x}) = \frac{\left\| |A^{-1}| \boldsymbol{f} + \sum_{j,k=1}^{n} \left| A^{-1} \frac{\partial A}{\partial a_{jk}} \boldsymbol{x} \right| e_{jk} \right\|_\infty}{\|\boldsymbol{x}\|_\infty}.$$

Then, since it is obvious that $\partial A / \partial a_{jk} = \boldsymbol{e}_j \boldsymbol{e}_k^T$, we have:

$$\begin{aligned} \sum_{j,k=1}^{n} \left| A^{-1} \frac{\partial A}{\partial a_{jk}} \boldsymbol{x} \right| e_{jk} &= \sum_{j,k=1}^{n} \left|A^{-1}(:,j)\right| |x_k| \, e_{jk} = \sum_{k=1}^{n} \left( \sum_{j=1}^{n} \left|A^{-1}(:,j)\right| e_{jk} \right) |x_k| \\ &= \sum_{k=1}^{n} \left|A^{-1}\right| E(:,k) \, |x_k| = \left|A^{-1}\right| \sum_{k=1}^{n} E(:,k) \, |x_k| = \left|A^{-1}\right| E \, |\boldsymbol{x}| , \end{aligned}$$

and the proof follows trivially. □

## 3.2 Basics on eigenvalue condition numbers

In this section we will present some well-known and some not so well-known results about eigenvalue condition numbers that are fundamental in Chapters 5 and 6 of this dissertation. In particular, in Section 3.2.1, we recall some standard results and definitions that are well described in the literature. On the other hand, Section 3.2.2 is devoted to the more general eigenvalue condition numbers for parameterizations, and it includes some new results for their computation.

We will only consider simple eigenvalues since for a simple eigenvalue $\lambda$ with left and right eigenvectors $\boldsymbol{y}$ and $\boldsymbol{x}$ respectively, we have $\boldsymbol{y}^* \boldsymbol{x} \neq 0$.

### 3.2.1 Standard results

Theorem 3.11 and its Corollary 3.12 can be found in [56]. They are the fundamental results from which all the other results in this section are derived.

The results in this section are valid for complex matrices. Note that any perturbation of a matrix $M \in \mathbb{C}^{n\times n}$ can be expressed as the sum $M + \delta M$, where $\delta M \in \mathbb{C}^{n\times n}$ is called the *perturbation matrix*.

**Theorem 3.11.** *Let $\lambda$ be a simple eigenvalue of $M \in \mathbb{C}^{n\times n}$, with left and right eigenvectors $\boldsymbol{y}$ and $\boldsymbol{x}$, respectively. Then, for any perturbation $\delta M \in \mathbb{C}^{n\times n}$ of $M$, there is a unique eigenvalue $\tilde{\lambda}$ of $M + \delta M$ such that*

$$\tilde{\lambda} = \lambda + \frac{\boldsymbol{y}^*(\delta M)\boldsymbol{x}}{\boldsymbol{y}^*\boldsymbol{x}} + \mathcal{O}(\|\delta M\|^2), \tag{3.4}$$

*where $\|\delta M\|$ is any norm of $\delta M$.*

**Corollary 3.12.** *Let $\lambda$ be a simple eigenvalue of $M \in \mathbb{C}^{n\times n}$ with left and right eigenvectors $\boldsymbol{y} = (y_1, \ldots, y_n)^T$ and $\boldsymbol{x} = (x_1, \ldots, x_n)^T$, respectively. Then $\lambda$ is a differentiable function of the entries $m_{ij}$ of $M$. Moreover,*

$$\frac{\partial \lambda}{\partial m_{ij}}(M) = \frac{\overline{y}_i x_j}{\boldsymbol{y}^*\boldsymbol{x}}.$$

On the other hand, in 1965, in [65], Wilkinson defined the notion of a condition number for simple eigenvalues. In the modern notation used, for instance, in [42], the Wilkinson condition number is defined as in Definition 3.13.

**Definition 3.13.** *Let $\lambda$ be a simple eigenvalue of $M \in \mathbb{C}^{n\times n}$. Then the* Wilkinson condition number *of $\lambda$, denoted by $\kappa_\lambda$, is defined as*

$$\kappa_\lambda := \lim_{\eta\to 0} \sup \left\{ \frac{|\delta\lambda|}{\eta} \,:\, (\lambda + \delta\lambda) \text{ is an eigenvalue of } (M + \delta M),\ \|\delta M\|_2 \le \eta \right\}.$$

Based on this definition, if a left eigenvector and a right eigenvector of a simple eigenvalue $\lambda$ of $M \in \mathbb{C}^{n\times n}$ are known, it is easy to compute the Wilkinson condition number of $\lambda$ [42] as it is proved in the next theorem.

**Theorem 3.14.** *Let $\lambda$ be a simple eigenvalue of $M \in \mathbb{C}^{n\times n}$, with left eigenvector $\boldsymbol{y} \in \mathbb{C}^n$ and right eigenvector $\boldsymbol{x} \in \mathbb{C}^n$. Then*

$$\kappa_\lambda = \frac{\|\boldsymbol{y}\|_2\|\boldsymbol{x}\|_2}{|\boldsymbol{y}^*\boldsymbol{x}|} = \frac{1}{\cos\angle(\boldsymbol{y}, \boldsymbol{x})}. \tag{3.5}$$

*Proof.* Observe that we can rewrite equation (3.4) as

$$\delta\lambda = \frac{\boldsymbol{y}^*\delta M\boldsymbol{x}}{\boldsymbol{y}^*\boldsymbol{x}} + \mathcal{O}(\|\delta M\|_2^2), \tag{3.6}$$

where we have considered the 2-norm, since this is the norm used in Definition 3.13 for $\kappa_\lambda$. Then, from Definition 3.13, we get

$$|\delta\lambda| \leq \frac{\|\boldsymbol{y}\|_2\|\boldsymbol{x}\|_2}{|\boldsymbol{y}^*\boldsymbol{x}|}\eta + \mathcal{O}(\eta^2) \Rightarrow \frac{|\delta\lambda|}{\eta} \leq \frac{\|\boldsymbol{y}\|_2\|\boldsymbol{x}\|_2}{|\boldsymbol{y}^*\boldsymbol{x}|} + \mathcal{O}(\eta) \Rightarrow \kappa_\lambda \leq \frac{\|\boldsymbol{y}\|_2\|\boldsymbol{x}\|_2}{|\boldsymbol{y}^*\boldsymbol{x}|}. \tag{3.7}$$

In order to prove the first equality in (3.5) we consider

$$\delta M = \frac{\eta}{\|\boldsymbol{y}\|_2\|\boldsymbol{x}\|_2}\boldsymbol{y}\boldsymbol{x}^*,$$

and note that

$$\|\delta M\|_2 = \frac{\eta}{\|\boldsymbol{y}\|_2\|\boldsymbol{x}\|_2}\|\boldsymbol{y}\|_2\|\boldsymbol{x}\|_2 = \eta.$$

Then, from (3.6), we have for this particular $\delta M$

$$\delta\lambda = \frac{\eta\|\boldsymbol{y}\|_2^2\|\boldsymbol{x}\|_2^2}{\|\boldsymbol{x}\|_2\|\boldsymbol{y}\|_2(\boldsymbol{y}^*\boldsymbol{x})} + \mathcal{O}(\eta^2) = \frac{\eta\|\boldsymbol{y}\|_2\|\boldsymbol{x}\|_2}{\boldsymbol{y}^*\boldsymbol{x}} + \mathcal{O}(\eta^2),$$

from where we obtain

$$\left|\frac{\delta\lambda}{\eta}\right| = \frac{\|\boldsymbol{y}\|_2\|\boldsymbol{x}\|_2}{|\boldsymbol{y}^*\boldsymbol{x}|} + \mathcal{O}(\eta).$$

Finally, by making $\eta$ to tend to 0 in the expression above and from the definition of the Wilkinson condition number we get

$$\kappa_\lambda \geq \frac{\|\boldsymbol{y}\|_2\|\boldsymbol{x}\|_2}{|\boldsymbol{y}^*\boldsymbol{x}|}. \tag{3.8}$$

The desired result follows from (3.7) and (3.8), since the last equality in (3.5) is a very well known fact from basic linear algebra. □

It is obvious, from its definition, that the Wilkinson condition number is an absolute-absolute normwise condition number, this means that it measures the absolute sensitivity of a simple eigenvalue with respect to absolute normwise perturbations of the matrix. In [10] the standard Wilkinson condition number was replaced by a relative-relative condition number, which measures a relative variation of the eigenvalues with respect to relative normwise perturbations of the matrix.

**Definition 3.15.** *Let $\lambda \neq 0$ be a simple eigenvalue of $M \in \mathbb{C}^{n\times n}$. Then we denote by $\kappa_\lambda^{\text{rel}}$ the* relative Wilkinson condition number of $\lambda$ *defined as*

$$\kappa_\lambda^{\text{rel}} := \lim_{\eta\to 0}\sup\left\{\frac{|\delta\lambda|}{\eta|\lambda|} : (\lambda+\delta\lambda) \text{ is an eigenvalue of } (M+\delta M),\ \|\delta M\|_2 \leq \eta\|M\|_2\right\}.$$

As before, this relative condition number can also be computed using a nice expression as stated in Theorem 3.16. This theorem can be proved in an almost identical way to Theorem 3.14 but, in this thesis, we prefer to use an alternative proof based on using the definitions of the condition numbers and the previous result for computing the Wilkinson condition number.

**Theorem 3.16.** *Let $\lambda \neq 0$ be a simple eigenvalue of $M \in \mathbb{C}^{n\times n}$ with left eigenvector $\boldsymbol{y} \in \mathbb{C}^n$ and right eigenvector $\boldsymbol{x} \in \mathbb{C}^n$. Then*

$$\kappa_\lambda^{\text{rel}} = \frac{\|\boldsymbol{y}\|_2\|\boldsymbol{x}\|_2}{|\boldsymbol{y}^*\boldsymbol{x}|}\,\frac{\|M\|_2}{|\lambda|} = \kappa_\lambda\,\frac{\|M\|_2}{|\lambda|}.$$

*Proof.* Note that by considering the change of variable $\eta' = \eta\|M\|_2$ on the expression in Definition 3.15 for the relative Wilkinson condition number, we have

$$\begin{aligned}\kappa_\lambda^{\text{rel}} &= \lim_{\eta'\to 0}\sup\left\{\frac{\|M\|_2|\delta\lambda|}{\eta'|\lambda|} : (\lambda+\delta\lambda) \text{ is an eigenvalue of } (M+\delta M),\ \|\delta M\|_2 \le \eta'\right\}\\ &= \frac{\|M\|_2}{|\lambda|}\lim_{\eta'\to 0}\sup\left\{\frac{|\delta\lambda|}{\eta'} : (\lambda+\delta\lambda) \text{ is an eigenvalue of } (M+\delta M),\ \|\delta M\|_2 \le \eta'\right\}\\ &= \kappa_\lambda\,\frac{\|M\|_2}{|\lambda|}.\end{aligned}$$

□

Following the ideas in [32], next we introduce a relative-relative componentwise condition number, that is, a measure of the relative variation of an eigenvalue with respect to the largest relative perturbation of each of the nonzero entries of the matrix. As in Section 3.1.1, we denote by $|M| \in \mathbb{C}^{n\times n}$ the matrix whose entries are the absolute values of the entries of $M$ (i.e., $|M|_{ij} := |M_{ij}|$) and we adopt a similar notation for vectors.

**Definition 3.17.** *Let $\lambda \neq 0$ be a simple eigenvalue of $M \in \mathbb{C}^{n\times n}$. We define the* relative componentwise condition number of $\lambda$ *as*

$$\operatorname{cond}(\lambda; M) := \lim_{\eta\to 0}\sup\left\{\frac{|\delta\lambda|}{\eta|\lambda|} : (\lambda+\delta\lambda) \textit{ is an eigenvalue of } (M+\delta M),\ |\delta M| \le \eta|M|\right\}.$$

The next theorem, stated for the first time in [37], gives an expression for computing $\operatorname{cond}(\lambda; M)$ and it can be seen as a consequence of the more general Theorem 3.23 we prove in Section 3.2.2, so we present its proof after Theorem 3.23.

**Theorem 3.18.** *Let $\lambda \neq 0$ be a simple eigenvalue with left eigenvector $\boldsymbol{y}$ and right eigenvector $\boldsymbol{x}$ of the matrix $M \in \mathbb{C}^{n\times n}$. Then*

$$\operatorname{cond}(\lambda; M) = \frac{|\boldsymbol{y}^*\|M\|\boldsymbol{x}|}{|\lambda\|\boldsymbol{y}^*\boldsymbol{x}|}. \tag{3.9}$$

A useful property that is easy to prove about this condition number is that

$$\operatorname{cond}(\lambda; M) \leq \sqrt{n}\,\kappa_\lambda^{\mathrm{rel}},$$

and, in many important situations, $\operatorname{cond}(\lambda; M)$ can be much smaller than $\kappa_\lambda^{\mathrm{rel}}$. For example, if we consider an entrywise positive matrix $P$, then there exists a positive eigenvalue $r$ of $P$, known as the *Perron-root*, which is strictly greater in absolute value than any other eigenvalue in the spectrum of $P$. Moreover, there exist entrywise positive left and right eigenvectors $\boldsymbol{y}_r$ and $\boldsymbol{x}_r$ of $\lambda$. Thus, it is obvious that

$$\operatorname{cond}(\lambda; P) = \frac{\boldsymbol{y}_r^* P \boldsymbol{x}_r}{\lambda \boldsymbol{y}_r^* \boldsymbol{x}_r} = 1,$$

meanwhile, $\kappa_\lambda^{\mathrm{rel}}$ can be arbitrary large [30].

Another important fact about $\operatorname{cond}(\lambda; M)$ is that it is invariant under diagonal similarity while Wilkinson and relative Wilkinson condition numbers are not.

**Lemma 3.19.** *For any matrix $K$ invertible and diagonal,*

$$\operatorname{cond}(\lambda; KMK^{-1}) = \operatorname{cond}(\lambda; M).$$

*Proof.* Let $G = KMK^{-1}$. Note that if $\boldsymbol{y}$ and $\boldsymbol{x}$ are left and right eigenvectors of the matrix $M$ associated to the simple eigenvalue $\lambda$, then $\boldsymbol{y}_K^* = \boldsymbol{y}^*K^{-1}$ and $\boldsymbol{x}_K = K\boldsymbol{x}$ are the corresponding left and right eigenvectors of $G$ associated to $\lambda$. Furthermore, since $K$ is diagonal, no addition occurs in $KMK^{-1}$ and we have that $|KMK^{-1}| = |K||M||K^{-1}|$. Consequently, $|\boldsymbol{y}_K^*\boldsymbol{x}_K| = |\boldsymbol{y}^*K^{-1}K\boldsymbol{x}| = |\boldsymbol{y}^*\boldsymbol{x}|$, and

$$\operatorname{cond}(\lambda; G) = \frac{|\boldsymbol{y}_K^*||G||\boldsymbol{x}_K|}{|\lambda||\boldsymbol{y}_K^*\boldsymbol{x}_K|} = \frac{|\boldsymbol{y}_K^*||KMK^{-1}||\boldsymbol{x}_K|}{|\lambda||\boldsymbol{y}^*\boldsymbol{x}|} = \frac{|\boldsymbol{y}^*||M||\boldsymbol{x}|}{|\lambda||\boldsymbol{y}^*\boldsymbol{x}|} = \operatorname{cond}(\lambda; M).$$

□

### 3.2.2 Eigenvalue condition numbers for parameterized matrices

As we have already commented, many interesting classes of matrices can be represented by sets of parameters different from its entries, whenever the entries are functions of certain parameters. Widely known examples include Cauchy, Vandermonde, and Toeplitz matrices [38, 41], among many others, and also the quasiseparable matrices considered in this thesis [24, 61]. This motivates us to extend the definitions in the previous section to more general representations and to focus on relative componentwise eigenvalue condition numbers for representations.

**Definition 3.20.** *Let $M \in \mathbb{C}^{n\times n}$ be a matrix whose entries are differentiable functions of a vector of parameters $\Omega = (\omega_1, \omega_2, \ldots, \omega_N)^T \in \mathbb{C}^N$. This is denoted by $M(\Omega)$. Let $\lambda \neq 0$ be a simple eigenvalue of $M(\Omega)$ with left eigenvector $\boldsymbol{y}$ and right eigenvector $\boldsymbol{x}$. Then define*

$$\operatorname{cond}(\lambda, M; \Omega) := \lim_{\eta\to 0} \sup \left\{ \frac{|\delta\lambda|}{\eta|\lambda|} : (\lambda + \delta\lambda) \text{ is an eigenvalue of } M(\Omega + \delta\Omega), \right.$$
$$\left. |\delta\Omega| \leq \eta|\Omega| \right\}.$$

*If the matrix $M$ is clear from the context, then we will usually denote by* $\operatorname{cond}(\lambda; \Omega)$ *the condition number* $\operatorname{cond}(\lambda, M; \Omega)$.

In order to find an explicit formula that allows us to calculate $\operatorname{cond}(\lambda; \Omega)$, as it can be done with $\operatorname{cond}(\lambda; M)$, the next definitions are convenient.

**Definition 3.21.** *Let $M \in \mathbb{C}^{n\times n}$ be a matrix whose entries are differentiable functions of a vector of parameters $\Omega = (\omega_1, \omega_2, \ldots, \omega_N)^T \in \mathbb{C}^N$. Let $\lambda \neq 0$ be a simple eigenvalue of $M(\Omega)$ with left eigenvector $\boldsymbol{y}$ and right eigenvector $\boldsymbol{x}$. We define the* relative gradient *of $\lambda$ with respect to $\Omega$ as the vector:*

$$\operatorname{relgrad}_\Omega(\lambda) := \left( \frac{\omega_1}{\lambda}\frac{\partial\lambda}{\partial\omega_1}, \ldots, \frac{\omega_N}{\lambda}\frac{\partial\lambda}{\partial\omega_N} \right)^T$$

*and the* relative perturbation of $\Omega$ *as the vector*

$$\operatorname{rel}\delta\Omega := \left( \frac{\delta\omega_1}{\omega_1}, \ldots, \frac{\delta\omega_N}{\omega_N} \right)^T,$$

*where if $w_i = 0$ for some $i$, then we define $\delta w_i / w_i \equiv 0$ in agreement with Definition 3.20.*

Taking into account our goals, the main result of this section is given in Theorem 3.23. But for proving that theorem, we will need the next proposition, which will also play an important role by itself in calculating the componentwise relative eigenvalue condition number for quasiseparable matrices with respect to parameters.

**Proposition 3.22.** *Let $M \in \mathbb{C}^{n\times n}$ be a matrix whose entries are differentiable functions of a vector of parameters $\Omega = (\omega_1, \omega_2, \ldots, \omega_N)^T \in \mathbb{C}^N$. This is denoted by $M(\Omega)$. Let $\lambda$ be a simple eigenvalue of $M(\Omega)$ with left eigenvector $\boldsymbol{y}$ and right eigenvector $\boldsymbol{x}$. Then*

$$\frac{\partial\lambda}{\partial\omega_i} = \frac{1}{\boldsymbol{y}^*\boldsymbol{x}} \left( \boldsymbol{y}^* \frac{\partial M(\Omega)}{\partial\omega_i} \boldsymbol{x} \right), \quad i \in \{1, ..., N\}. \tag{3.10}$$

*Proof.* Compute explicitly the partial derivative of $M(\Omega)\boldsymbol{x} = \lambda\boldsymbol{x}$ to get

$$\frac{\partial M(\Omega)}{\partial \omega_i}\boldsymbol{x} + M(\Omega)\frac{\partial \boldsymbol{x}}{\partial \omega_i} = \frac{\partial \lambda}{\partial \omega_i}\boldsymbol{x} + \lambda\frac{\partial \boldsymbol{x}}{\partial \omega_i},$$

then, we multiply on the left the last equation by $\boldsymbol{y}^*$ and cancel out equal terms to find

$$\boldsymbol{y}^*\frac{\partial M(\Omega)}{\partial \omega_i}\boldsymbol{x} = \frac{\partial \lambda}{\partial \omega_i}\boldsymbol{y}^*\boldsymbol{x},$$

which completes the proof. □

**Theorem 3.23.** *Under the same hypotheses of Definition 3.20:*

$$\operatorname{cond}(\lambda;\Omega) = \sum_{i=1}^{N}\left|\frac{\omega_i}{\lambda}\frac{\partial \lambda}{\partial \omega_i}\right| = \|\operatorname{relgrad}_\Omega(\lambda)\|_1 \tag{3.11}$$

*and*

$$\frac{\omega_i}{\lambda}\frac{\partial \lambda}{\partial \omega_i} = \frac{1}{\lambda(\boldsymbol{y}^*\boldsymbol{x})}\boldsymbol{y}^*\left(\omega_i\frac{\partial M(\Omega)}{\partial \omega_i}\right)\boldsymbol{x}, \quad \textit{for } i = 1,\ldots,N. \tag{3.12}$$

*Proof.* We can form the absolute gradient vector by considering all the partial derivatives of $\lambda$ with respect to $\omega_i$:

$$\operatorname{grad}_\Omega(\lambda) = \left(\frac{\partial \lambda}{\partial \omega_1},\ldots,\frac{\partial \lambda}{\partial \omega_k},\ldots,\frac{\partial \lambda}{\partial \omega_N}\right)^T,$$

and, for infinitesimal absolute perturbations, $\delta\Omega := (\delta\omega_1,\ldots,\delta\omega_k,\ldots,\delta\omega_N)^T$, we have

$$\delta\lambda = \operatorname{grad}_\Omega(\lambda)^T\cdot\delta\Omega + \text{ higher order terms (h.o.t)}. \tag{3.13}$$

Since, according to Definition 3.20, $\delta\omega_i = 0$ whenever $\omega_i = 0$, following the convention in Definition 3.21, we rewrite (3.13) as

$$\frac{\delta\lambda}{\lambda} = \left(\frac{\omega_1}{\lambda}\frac{\partial \lambda}{\partial \omega_1},\ldots,\frac{\omega_N}{\lambda}\frac{\partial \lambda}{\partial \omega_N}\right)\cdot\left(\frac{\delta\omega_1}{\omega_1},\ldots,\frac{\delta\omega_N}{\omega_N}\right)^T + (\text{h.o.t}), \tag{3.14}$$

which can be rewritten in the notation from Definition 3.21 as

$$\frac{\delta\lambda}{\lambda} = \operatorname{relgrad}_\Omega(\lambda)^T\cdot\operatorname{rel}\delta\Omega + (\text{h.o.t}). \tag{3.15}$$

Note that $|\delta\Omega| \le \eta|\Omega|$ implies $|\operatorname{rel}\delta\Omega| \le \eta(1,1,\ldots,1)^T$, where $0 < \eta \ll 1$. Thus, $||\operatorname{rel}\delta\Omega||_\infty \le \eta$, and by applying the Hölder inequality ($|u^Tv| \le ||u||_1||v||_\infty$) in equation (3.15), we obtain

$$\left|\frac{\delta\lambda}{\lambda}\right| \le ||\operatorname{relgrad}_\Omega(\lambda)||_1||\operatorname{rel}\delta\Omega||_\infty + (\text{h.o.t.}) \le \eta||\operatorname{relgrad}_\Omega(\lambda)||_1 + (\text{h.o.t.}). \tag{3.16}$$

From standard properties of norms (see [41, Ch. 6]), there exist particular vectors $\operatorname{rel}\delta\Omega$ with infinity norm $\eta$ such that $|\operatorname{relgrad}_\Omega(\lambda)^T \cdot \operatorname{rel}\delta\Omega| = \|\operatorname{relgrad}_\Omega(\lambda)\|_1 \|\operatorname{rel}\delta\Omega\|_\infty$. Hence, for these vectors $\operatorname{rel}\delta\Omega$, we have

$$\left|\frac{\delta\lambda}{\lambda}\right| = ||\operatorname{relgrad}_\Omega(\lambda)||_1 ||\operatorname{rel}\delta\Omega||_\infty + (\text{h.o.t.}) = \eta||\operatorname{relgrad}_\Omega(\lambda)||_1 + (\text{h.o.t.}). \quad (3.17)$$

From (3.16), (3.17) and Definition 3.20 we prove immediately that if $\lambda \neq 0$, then (3.11) holds. Equation (3.12) follows from Proposition 3.22. □

Next, we prove Theorem 3.18 as a particular case of Theorem 3.23, when the representation is given by the entries of $M$ themselves (i.e., $\Omega = (m_{ij})$, with the entries of $M$ written as a long vector). In this case,

$$m_{ij}\frac{\partial M}{\partial m_{ij}} = m_{ij}\boldsymbol{e}_i\boldsymbol{e}_j{}^T,$$

where $\boldsymbol{e}_i$ and $\boldsymbol{e}_j$ are the respective $i$-th and $j$th canonical vectors in $\mathbb{C}^n$. Then, we can rewrite equations (3.12) and (3.11), respectively, as

$$\frac{m_{ij}}{\lambda}\frac{\partial\lambda}{\partial m_{ij}} = \frac{1}{\lambda(\boldsymbol{y}^*\boldsymbol{x})}\bar{y}_i m_{ij} x_j,$$

$$\operatorname{cond}(\lambda; M) = \sum_{i=1,j=1}^{n} \frac{1}{|\lambda||(\boldsymbol{y}^*\boldsymbol{x})|}|\bar{y}_i||m_{ij}||x_j| = \frac{|\boldsymbol{y}^*||M||\boldsymbol{x}|}{|\lambda||\boldsymbol{y}^*\boldsymbol{x}|},$$

which is Theorem 3.18.

In this section, for practical purposes and in order to keep consistency with the results presented in [32], we have only considered, so far, relative perturbations with respect to the parameters in the representation $\Omega$. Nevertheless, if we consider relative perturbations on the parameters in $\Omega$ but with respect to a general vector of parameters $E$ we obtain the more general condition number in Definition 3.24 and the consequent expression for its computation given in Theorem 3.25.

**Definition 3.24.** *Let $M \in \mathbb{C}^{n\times n}$ be a matrix whose entries are differentiable functions of a vector of parameters $\Omega = (\omega_1, \omega_2, \ldots, \omega_N)^T \in \mathbb{C}^N$. This is denoted by $M(\Omega)$. Let $E = (e_1, e_2, \ldots, e_N)^T \in \mathbb{R}^N$ with nonnegative entries and $\lambda \neq 0$ be a simple eigenvalue of $M(\Omega)$ with left eigenvector $\boldsymbol{y}$ and right eigenvector $\boldsymbol{x}$. Then define*

$$\operatorname{cond}_E(\lambda, M; \Omega) := \lim_{\eta\to 0} \sup\left\{\frac{|\delta\lambda|}{\eta|\lambda|} : (\lambda+\delta\lambda) \text{ is an eigenvalue of } M(\Omega+\delta\Omega), \right.$$
$$\left. |\delta\Omega| \leq \eta E\right\}.$$

*If the matrix $M$ is clear from the context, then we will usually denote by $\operatorname{cond}_E(\lambda; \Omega)$ the condition number $\operatorname{cond}_E(\lambda, M; \Omega)$.*

**Theorem 3.25.** *Under the same hypotheses of Definition 3.24:*

$$\mathrm{cond}_E(\lambda;\Omega) = \frac{1}{|\lambda\,(\boldsymbol{y}^*\boldsymbol{x})|}\sum_{i=1}^{N}\left|\boldsymbol{y}^*\left(\frac{\partial M(\Omega)}{\partial\omega_i}\right)\boldsymbol{x}\right|e_i. \tag{3.18}$$

*Proof.* Since $\lambda$ is a differentiable function of the entries of the matrix $M$, and those entries are differentiable functions of the parameters in $\Omega$, we have that $\lambda$ is a differentiable function of the parameters in $\Omega$, and, using differential calculus, we have:

$$\delta\lambda = \sum_{i=1}^{N}\frac{\partial\lambda}{\partial\omega_i}\delta\omega_i + \text{ higher order terms (h.o.t)},$$

and therefore:

$$\frac{\delta\lambda}{\lambda} = \sum_{i=1}^{N}\left(\frac{\omega_i}{\lambda}\frac{\partial\lambda}{\partial\omega_i}\right)\frac{\delta\omega_i}{\omega_i} + (\text{h.o.t}), \tag{3.19}$$

where even in the case $\omega_i = 0$ we consider formally $\dfrac{\omega_i}{\omega_i} = 1$.

From (3.19), using equation (3.12) from Theorem 3.23, standard properties of the absolute value, and the inequality $|\delta\Omega| \leq \eta E$, we get

$$\left|\frac{\delta\lambda}{\lambda}\right| \leq \eta\sum_{i=1}^{N}\left|\frac{1}{\lambda(\boldsymbol{y}^*\boldsymbol{x})}\boldsymbol{y}^*\left(\omega_i\frac{\partial M(\Omega)}{\partial\omega_i}\right)\boldsymbol{x}\right|\left|\frac{e_i}{\omega_i}\right| + (\text{h.o.t}). \tag{3.20}$$

Then, if $\eta$ tends to zero, from (3.20) and from Definition 3.24, it is straightforward to get

$$\mathrm{cond}_E(\lambda;\Omega) \leq \frac{1}{|\lambda\,(\boldsymbol{y}^*\boldsymbol{x})|}\sum_{i=1}^{N}\left|\boldsymbol{y}^*\left(\frac{\partial M(\Omega)}{\partial\omega_i}\right)\boldsymbol{x}\right|e_i. \tag{3.21}$$

On the other hand, if we consider the perturbations:

$$\delta\omega_i = \eta\left[\mathrm{sign}\left(\frac{1}{\lambda(\boldsymbol{y}^*\boldsymbol{x})}\boldsymbol{y}^*\left(\omega_i\frac{\partial M(\Omega)}{\partial\omega_i}\right)\boldsymbol{x}\right)\right]e_i, \quad \text{for } i = 1,\ldots,N,$$

we can obtain, from (3.19) and from Definition 3.24, the desired equality in (3.21). □

**Remark 3.26.** *It is obvious that Theorem 3.23 can be seen as a particular case of Theorem 3.25 when $E = \Omega$, but since this is the most natural, and important in practice, choice for $E$, and the proof of Theorem 3.25 has the same flavor of the proof of Theorem 3.10, we have stated and proved Theorem 3.23 independently in order to provide an alternative proof using properties of norms, scalar products, and the notation in Definition 3.21, which we consider useful for future tasks.*

# Chapter 4

# Structured condition numbers for linear systems with parameterized quasiseparable coefficient matrices

In this chapter we study condition numbers for the solution of a linear system of equations whose coefficient matrix is $\{1,1\}$-quasiseparable with respect to relative componentwise perturbations on the parameters in the quasiseparable and in the Givens-vector representations of the $\{1,1\}$-quasiseparable coefficient matrix. Sections 4.1, 4.2, 4.3 and 4.4 include our original contributions for these condition numbers. In particular, in Sections 4.1 and 4.2 we provide expressions for the respective condition numbers in the quasiseparable and the Givens-vector representations (see Theorems 4.2 and 4.11 respectively), and prove in Proposition 4.5 that the structured condition number in the quasiseparable representation can not be much larger than the unstructured standard one but it can be much smaller as we have observed in our numerical experiments described in Section 4.5. In Section 4.3, we compare these condition numbers and prove that the Givens-vector representation is a more stable representation (see Theorem 4.13) for the solution of linear systems since produces smaller condition numbers, but that the differences between the respective structured condition numbers can not be too large (see Theorem 4.16) and therefore both representations can be considered numerically equivalent. In Section 4.4, it is shown how to estimate both condition numbers via an effective condition number which can be computed fast, i.e., in $\mathcal{O}(n)$ operations, as proved in Proposition 4.20.

## 4.1 Condition number of the solution of $\{1,1\}$-quasiseparable linear systems in the quasiseparable representation

Taking into account that our goal is to obtain explicit expressions of structured condition numbers for the solution of a linear system involving a quasiseparable matrix in the quasiseparable representation by using differential calculus via Theorem 3.10, the next lemma will become useful.

**Lemma 4.1.** *Let $A \in \mathbb{R}^{n\times n}$ be a $\{1,1\}-$quasiseparable matrix and $A = A_L + A_D + A_U$, with $A_L$ strictly lower triangular, $A_D$ diagonal, and $A_U$ strictly upper triangular. Let $\Omega_{QS}$ be a quasiseparable representation of $A$, where $\Omega_{QS} = (\{p_i\}_{i=2}^n, \{a_i\}_{i=2}^{n-1}, \{q_i\}_{i=1}^{n-1}, \{d_i\}_{i=1}^n, \{g_i\}_{i=1}^{n-1}, \{b_i\}_{i=2}^{n-1}, \{h_i\}_{i=2}^n)$. Then the entries of $A$ are differentiable functions of the parameters in $\Omega_{QS}$ and:*

*a)* $\dfrac{\partial A}{\partial d_i} = \boldsymbol{e}_i \boldsymbol{e}_i^T$, *for* $i = 1, \ldots, n$.

*b)* $p_i \dfrac{\partial A}{\partial p_i} = \boldsymbol{e}_i A_L(i,:)$, *for* $i = 2, \ldots, n$.

*c)* $a_i \dfrac{\partial A}{\partial a_i} = \begin{bmatrix} 0 & 0 \\ A(i+1:n, 1:i-1) & 0 \end{bmatrix}$, *for* $i = 2, \ldots, n-1$.

*d)* $q_i \dfrac{\partial A}{\partial q_i} = A_L(:,i)\boldsymbol{e}_i^T$, *for* $i = 1, \ldots, n-1$.

*e)* $g_i \dfrac{\partial A}{\partial g_i} = \boldsymbol{e}_i A_U(i,:)$, *for* $i = 1, \ldots, n-1$.

*f)* $b_i \dfrac{\partial A}{\partial b_i} = \begin{bmatrix} 0 & A(1:i-1, i+1:n) \\ 0 & 0 \end{bmatrix}$, *for* $i = 2, \ldots, n-1$.

*g)* $h_i \dfrac{\partial A}{\partial h_i} = A_U(:,i)\boldsymbol{e}_i^T$, *for* $i = 2, \ldots, n$,

*where $\boldsymbol{e}_i$, for $i = 1, \ldots, n$, denotes the $i$-th canonical vector in the basis of $\mathbb{R}^n$.*

*Proof.* That the entries of $A$ are differentiable with respect to the parameters in $\Omega_{QS}$ it is straightforward from the explicit expression given in Theorem 2.18 for the entries of $A$ in terms of such parameters. Furthermore, the expressions in a), b), c), d), e), f), and g), are easily obtained by considering the corresponding partial derivatives in the expression for $A$ in Theorem 2.18. □

Since from Lemma 4.1 we have that the entries of a quasiseparable matrix $A$ are differentiable functions of the parameters in a quasiseparable representation of $A$, we can deduce relative-relative componentwise condition numbers of the solution of linear systems with respect to these representations by using Theorem 3.10. This leads to Theorem 4.2.

**Theorem 4.2.** *Let $A\boldsymbol{x} = \boldsymbol{b}$, where $0 \neq \boldsymbol{x} \in \mathbb{R}^n$, $A \in \mathbb{R}^{n\times n}$ is a nonsingular $\{1,1\}$-quasiseparable matrix with a quasiseparable representation $\Omega_{QS}$, and such that $A = A_L + A_D + A_U$, with $A_L$ strictly lower triangular, $A_D$ diagonal, and $A_U$ strictly upper triangular. Let $0 \leq \boldsymbol{f} \in \mathbb{R}^n$ and $0 \leq E_{QS} \in \mathbb{R}^{7n-8}$. Then*

$$\begin{aligned}\operatorname{cond}_{E_{QS},\boldsymbol{f}}(A(\Omega_{QS}),\boldsymbol{x}) =& \frac{1}{\|\boldsymbol{x}\|_\infty}\Bigg\| \left|A^{-1}\right| \boldsymbol{f} + |A^{-1}|Q_d|\boldsymbol{x}| + |A^{-1}||Q_p||A_L\boldsymbol{x}| \\ &+ |A^{-1}A_L||Q_q||\boldsymbol{x}| + |A^{-1}||Q_g||A_U\boldsymbol{x}| + |A^{-1}A_U||Q_h||\boldsymbol{x}| \\ &+ \sum_{i=2}^{n-1}\left|A^{-1}\begin{bmatrix} 0 & 0 \\ A(i+1:n,1:i-1) & 0\end{bmatrix}\boldsymbol{x}\right|\left|\frac{e_{a_i}}{a_i}\right| \\ &+ \sum_{j=2}^{n-1}\left|A^{-1}\begin{bmatrix} 0 & A(1:j-1,j+1:n) \\ 0 & 0\end{bmatrix}\boldsymbol{x}\right|\left|\frac{e_{b_i}}{b_i}\right|\Bigg\|_\infty,\end{aligned}$$

*where:*

$$\begin{aligned}\Omega_{QS} &= (\{p_i\}_{i=2}^n, \{a_i\}_{i=2}^{n-1}, \{q_i\}_{i=1}^{n-1}, \{d_i\}_{i=1}^n, \{g_i\}_{i=1}^{n-1}, \{b_i\}_{i=2}^{n-1}, \{h_i\}_{i=2}^n), \\ E_{QS} &= (\{e_{p_i}\}_{i=2}^n, \{e_{a_i}\}_{i=2}^{n-1}, \{e_{q_i}\}_{i=1}^{n-1}, \{e_{d_i}\}_{i=1}^n, \{e_{g_i}\}_{i=1}^{n-1}, \{e_{b_i}\}_{i=2}^{n-1}, \{e_{h_i}\}_{i=2}^n), \\ Q_d &= \operatorname{diag}(e_{d_1},\ldots,e_{d_n}),\ Q_p = \operatorname{diag}\left(1, \frac{e_{p_2}}{p_2},\ldots,\frac{e_{p_n}}{p_n}\right), \\ Q_q &= \operatorname{diag}\left(\frac{e_{q_1}}{q_1},\ldots,\frac{e_{q_{n-1}}}{q_{n-1}},1\right),\ Q_g = \operatorname{diag}\left(\frac{e_{g_1}}{g_1},\ldots,\frac{e_{g_{n-1}}}{g_{n-1}},1\right), \\ Q_h &= \operatorname{diag}\left(1, \frac{e_{h_2}}{h_2},\ldots,\frac{e_{h_n}}{h_n}\right),\end{aligned}$$

*and each quotient whose denominator is zero must be understood as zero if the numerator is also zero and, otherwise, the zero parameter in the denominator should be formally cancelled out with the same parameter in the corresponding piece of $A$.*

*Proof.* We will proceed by calculating the contribution of each subset of parameters to the expression for $\operatorname{cond}_{E_{QS},\boldsymbol{f}}(A(\Omega_{QS}),\boldsymbol{x})$ given in Theorem 3.10 as follows.
Derivatives with respect to $\{d_i\}_{i=1}^n$. By using a) in Lemma 4.1 we get:

$$\kappa_d := \sum_{i=1}^n \left|A^{-1}\frac{\partial A}{\partial d_i}\boldsymbol{x}\right| e_{d_i} = \sum_{i=1}^n \left|A^{-1}\boldsymbol{e}_i\boldsymbol{e}_i^T\boldsymbol{x}\right| e_{d_i} = \sum_{i=1}^n \left|A^{-1}(:,i)\right| |x_i|\, e_{d_i} = \left|A^{-1}\right| |Q_d|\, |\boldsymbol{x}|\,.$$

Derivatives with respect to $\{p_i\}_{i=2}^n$. From b) in Lemma 4.1 we have:

$$\kappa_p := \sum_{i=2}^{n} \left| A^{-1} \frac{\partial A}{\partial p_i} \boldsymbol{x} \right| |e_{p_i}| = \sum_{i=2}^{n} \left| A^{-1} p_i \frac{\partial A}{\partial p_i} \boldsymbol{x} \right| \left| \frac{e_{p_i}}{p_i} \right| = \sum_{i=2}^{n} \left| A^{-1} \boldsymbol{e}_i A_L(i,:) \boldsymbol{x} \right| \left| \frac{e_{p_i}}{p_i} \right|$$
$$= \sum_{i=2}^{n} \left| A^{-1}(:,i) A_L(i,:) \boldsymbol{x} \right| \left| \frac{e_{p_i}}{p_i} \right| = \sum_{i=2}^{n} \left| A^{-1}(:,i) \right| \left| \frac{e_{p_i}}{p_i} \right| |A_L(i,:)\boldsymbol{x}| = \left| A^{-1} \right| |Q_p| \, |A_L \boldsymbol{x}| \, .$$

Derivatives with respect to $\{a_i\}_{i=2}^{n-1}$. From c) in Lemma 4.1 we have:

$$\kappa_a := \sum_{i=2}^{n-1} \left| A^{-1} a_i \frac{\partial A}{\partial a_i} \boldsymbol{x} \right| \left| \frac{e_{a_i}}{a_i} \right| = \sum_{i=2}^{n-1} \left| A^{-1} \begin{bmatrix} 0 & 0 \\ A(i+1:n, 1:i-1) & 0 \end{bmatrix} \boldsymbol{x} \right| \left| \frac{e_{a_i}}{a_i} \right|.$$

Derivatives with respect to $\{q_j\}_{j=1}^{n-1}$. From d) in Lemma 4.1 we have:

$$\kappa_q := \sum_{j=1}^{n-1} \left| A^{-1} \frac{\partial A}{\partial q_j} \boldsymbol{x} \right| \left| e_{q_j} \right| = \sum_{j=1}^{n-1} \left| A^{-1} q_j \frac{\partial A}{\partial q_j} \boldsymbol{x} \right| \left| \frac{e_{q_j}}{q_j} \right| = \sum_{j=1}^{n-1} \left| A^{-1} A_L(:,j) \boldsymbol{e}_j^T \boldsymbol{x} \right| \left| \frac{e_{q_j}}{q_j} \right|$$
$$= \sum_{j=1}^{n-1} \left| A^{-1} A_L(:,j) \right| |x_j| \left| \frac{e_{q_j}}{q_j} \right| = \left| A^{-1} A_L \right| |Q_q| \, |\boldsymbol{x}| \, .$$

Analogously, we can find the contribution to $\mathrm{cond}_{E_{QS},\boldsymbol{f}}(A(\Omega_{QS}), \boldsymbol{x})$ of the derivatives of $A$ with respect to the parameters $\{g_i\}_{i=1}^{n-1}, \{b_i\}_{i=2}^{n-1}$, and $\{h_i\}_{i=2}^{n}$, which describe the strictly upper triangular part of $A$. The results are the following.

Derivatives with respect to $\{g_i\}_{i=1}^{n-1}$. By using e) in Lemma 4.1 we obtain:

$$\kappa_g := \sum_{i=1}^{n-1} \left| A^{-1} \frac{\partial A}{\partial g_i} \boldsymbol{x} \right| |e_{g_i}| = \left| A^{-1} \right| |Q_g| \, |A_U \boldsymbol{x}| \, .$$

Derivatives with respect to $\{b_i\}_{i=2}^{n-1}$. By using f) in Lemma 4.1 we obtain:

$$\kappa_b := \sum_{i=2}^{n-1} \left| A^{-1} \frac{\partial A}{\partial b_i} \boldsymbol{x} \right| |e_{b_i}| = \sum_{i=2}^{n-1} \left| A^{-1} \begin{bmatrix} 0 & A(1:i-1, i+1:n) \\ 0 & 0 \end{bmatrix} \boldsymbol{x} \right| \left| \frac{e_{b_i}}{b_i} \right| .$$

Derivatives with respect to $\{h_j\}_{j=2}^{n}$. By using g) in Lemma 4.1 we obtain:

$$\kappa_h := \sum_{j=2}^{n} \left| A^{-1} \frac{\partial A}{\partial h_j} \boldsymbol{x} \right| \left| e_{h_j} \right| = \left| A^{-1} A_U \right| |Q_h| \, |\boldsymbol{x}| \, .$$

This proof is completed by observing that according to Theorem 3.10 we have:

$$\mathrm{cond}_{E_{QS},\boldsymbol{f}}(A(\Omega_{QS}), \boldsymbol{x}) = \frac{\left\| |A^{-1}| \, \boldsymbol{f} + \kappa_d + \kappa_p + \kappa_a + \kappa_q + \kappa_g + \kappa_b + \kappa_h \right\|_\infty}{\|\boldsymbol{x}\|_\infty}.$$

□

As it is easy to see from its expression in Theorem 4.2, $\mathrm{cond}_{E_{QS},\boldsymbol{f}}(A(\Omega_{QS}),\boldsymbol{x})$ depends in general on the quasiseparable representation $\Omega_{QS}$ of the matrix $A$. More specifically, that condition number depends on the ratios between the parameters in the representation and the corresponding tolerances in $E_{QS}$. Next, we will restrict ourselves to the case $E_{QS} = |\Omega_{QS}|$ in Theorem 4.3, which is the most natural election for $E_{QS}$. In this situation we adopt for brevity in the rest of this work, the following notation:

$$\mathrm{cond}_{\boldsymbol{f}}(A(\Omega_{QS}),\boldsymbol{x}) \equiv \mathrm{cond}_{|\Omega_{QS}|,\boldsymbol{f}}(A(\Omega_{QS}),\boldsymbol{x}),$$

since the parameters $\Omega_{QS}$ are already shown in $A(\Omega_{QS})$.

**Theorem 4.3.** *Let $A\boldsymbol{x} = \boldsymbol{b}$, where $0 \neq \boldsymbol{x} \in \mathbb{R}^n$ and $A \in \mathbb{R}^{n\times n}$ is a nonsingular $\{1,1\}$-quasiseparable matrix such that $A = A_L + A_D + A_U$, with $A_L$ strictly lower triangular, $A_D$ diagonal, and $A_U$ strictly upper triangular. Let $0 \leq \boldsymbol{f} \in \mathbb{R}^n$. Then*

$$\begin{aligned}\mathrm{cond}_{\boldsymbol{f}}(A(\Omega_{QS}),\boldsymbol{x}) = &\frac{1}{\|\boldsymbol{x}\|_\infty}\Bigg\| \left|A^{-1}\right| \boldsymbol{f} + |A^{-1}||A_D||\boldsymbol{x}| + |A^{-1}||A_L\boldsymbol{x}| \\ &+ |A^{-1}A_L||\boldsymbol{x}| + |A^{-1}||A_U\boldsymbol{x}| + |A^{-1}A_U||\boldsymbol{x}| \\ &+ \sum_{i=2}^{n-1} \left| A^{-1} \begin{bmatrix} 0 & 0 \\ A(i+1:n,1:i-1) & 0 \end{bmatrix} \boldsymbol{x} \right| \\ &+ \sum_{j=2}^{n-1} \left| A^{-1} \begin{bmatrix} 0 & A(1:j-1,j+1:n) \\ 0 & 0 \end{bmatrix} \boldsymbol{x} \right| \Bigg\|_\infty .\end{aligned}$$

*Proof.* It follows directly from the expression in Theorem 4.2 for $\mathrm{cond}_{E_{QS},\boldsymbol{f}}(A(\Omega_{QS}),\boldsymbol{x})$ by observing that, in this case, we are considering $E_{QS} = |\Omega_{QS}|$ and, therefore, using the notation in Theorem 4.2, the following equalities hold: $Q_d = |A_D|$, $|Q_p| = |Q_q| = |Q_g| = |Q_h| = I$, and $|e_{a_i}/a_i| = |e_{b_i}/b_i| = 1$. □

Proposition 4.4 proves that $\mathrm{cond}_{\boldsymbol{f}}(A(\Omega_{QS}),\boldsymbol{x})$ depends only on $A$, $\boldsymbol{x}$ and $\boldsymbol{f}$, but not on the particular choice of quasiseparable parameters.

**Proposition 4.4.** *Let $A\boldsymbol{x} = \boldsymbol{b}$, where $0 \neq \boldsymbol{x} \in \mathbb{R}^n$ and $A \in \mathbb{R}^{n\times n}$ is a nonsingular $\{1,1\}$-quasiseparable matrix. Let $0 \leq \boldsymbol{f} \in \mathbb{R}^n$. Then, for any two vectors $\Omega_{QS}$ and $\Omega'_{QS}$ of quasiseparable parameters of $A$,*

$$\mathrm{cond}_{\boldsymbol{f}}(A(\Omega_{QS}),\boldsymbol{x}) = \mathrm{cond}_{\boldsymbol{f}}(A(\Omega'_{QS}),\boldsymbol{x}).$$

*Proof.* It is obvious from the fact that the expression in Theorem 4.3 does not depend on the parameters of the representation $\Omega_{QS}$ but on the entries of the matrix $A$ and the entries of the vectors $\boldsymbol{x}$ and $\boldsymbol{f}$. □

Proposition 4.5 states another important property of this relative componentwise condition number that arises from the natural comparison with the unstructured relative entrywise condition number for the solution of linear systems defined in Definition 3.4 and further developed in Theorem 3.5.

**Proposition 4.5.** *Let $A\boldsymbol{x} = \boldsymbol{b}$, where $0 \neq \boldsymbol{x} \in \mathbb{R}^n$ and $A \in \mathbb{R}^{n\times n}$ is a nonsingular $\{1,1\}$-quasiseparable matrix, and let $\Omega_{QS}$ be a quasiseparable representation of $A$. Then, for $0 \leq \boldsymbol{f} \in \mathbb{R}^n$, the following relation holds,*

$$\operatorname{cond}_{\boldsymbol{f}}(A(\Omega_{QS}), \boldsymbol{x}) \leq n \operatorname{cond}_{|A|,\boldsymbol{f}}(A, \boldsymbol{x}).$$

*Proof.* From Theorem 4.3 and using standard properties of absolute values and norms we obtain:

$$\begin{aligned}
\operatorname{cond}_{\boldsymbol{f}}(A(\Omega_{QS}), \boldsymbol{x}) \leq & \frac{1}{\|\boldsymbol{x}\|_\infty} \Big\| \left|A^{-1}\right| \boldsymbol{f} + \left|A^{-1}\right| |A_D| \, |\boldsymbol{x}| + \left|A^{-1}\right| |A_L| \, |\boldsymbol{x}| \\
& + \left|A^{-1}\right| |A_L| \, |\boldsymbol{x}| + \left|A^{-1}\right| |A_U| \, |\boldsymbol{x}| + \left|A^{-1}\right| |A_U| \, |\boldsymbol{x}| \\
& + \sum_{i=2}^{n-1} \left|A^{-1}\right| |A_L| \, |\boldsymbol{x}| + \sum_{i=2}^{n-1} \left|A^{-1}\right| |A_U| \, |\boldsymbol{x}| \Big\|_\infty \\
= & \frac{1}{\|\boldsymbol{x}\|_\infty} \Big\| \left|A^{-1}\right| \boldsymbol{f} + \left|A^{-1}\right| |A_D| \, |\boldsymbol{x}| \\
& + n \left|A^{-1}\right| |A_L| \, |\boldsymbol{x}| + n \left|A^{-1}\right| |A_U| \, |\boldsymbol{x}| \Big\|_\infty \\
\leq & \frac{n}{\|\boldsymbol{x}\|_\infty} \Big\| \left|A^{-1}\right| \boldsymbol{f} + \left|A^{-1}\right| |A| \, |\boldsymbol{x}| \Big\|_\infty = n \operatorname{cond}_{|A|,\boldsymbol{f}}(A, \boldsymbol{x}).
\end{aligned}$$

□

According to this proposition, the structured condition number $\operatorname{cond}_{\boldsymbol{f}}(A(\Omega_{QS}), \boldsymbol{x})$ is smaller than the unstructured condition number $\operatorname{cond}_{|A|,\boldsymbol{f}}(A, \boldsymbol{x})$, except for a factor $n$. In addition, we will see in the numerical experiments presented in Section 4.5 that it can be much smaller.

From Proposition 3.6 we know that the unstructured componentwise condition number is invariant under row scaling, which is a very convenient property (see [41, Secs. 7.2 and 7.3]). Therefore, it makes sense to study the behavior of the structured condition number under row scaling for the natural choice $\boldsymbol{f} = |\boldsymbol{b}|$ as well. This is done in Proposition 4.7, for which we will need Lemma 4.6. Proposition 4.7 proves that the structured componentwise condition number is also invariant under row scaling.

**Lemma 4.6.** *Let $K = \operatorname{diag}(k_1, k_2, \cdots, k_n)$ be an invertible diagonal matrix and $A \in \mathbb{R}^{n\times n}$ be a $\{1,1\}$-quasiseparable matrix with a quasiseparable representation $\Omega_{QS} = \left(\{p_i\}_{i=2}^n, \{a_i\}_{i=2}^{n-1}, \{q_i\}_{i=1}^{n-1}, \ \{d_i\}_{i=1}^n, \{g_i\}_{i=1}^{n-1}, \{b_i\}_{i=2}^{n-1}, \{h_i\}_{i=2}^n\right)$, as in Theorem 2.18. Then, the matrix $KA$ is also a $\{1,1\}$-quasiseparable matrix and $\Omega'_{QS} = (\{k_i p_i\}_{i=2}^n, \{a_i\}_{i=2}^{n-1}, \{q_i\}_{i=1}^{n-1}, \{k_i d_i\}_{i=1}^n, \{k_i g_i\}_{i=1}^{n-1}, \{b_i\}_{i=2}^{n-1}, \{h_i\}_{i=2}^n)$ is a quasiseparable representation of $KA$.*

*Proof.* It follows from Theorem 2.18 and from $(KA)(i,j) = k_i A(i,j)$. □

**Proposition 4.7.** *Let $A\boldsymbol{x} = \boldsymbol{b}$, where $0 \neq \boldsymbol{x} \in \mathbb{R}^n$ and $A \in \mathbb{R}^{n\times n}$ is a nonsingular $\{1,1\}$-quasiseparable matrix with a quasiseparable representation $\Omega_{QS}$, such that $A = A_L + A_D + A_U$, with $A_L$ strictly lower triangular, $A_D$ diagonal, and $A_U$ strictly upper triangular. Let $K \in \mathbb{R}^{n\times n}$ be an invertible diagonal matrix. Then*

$$\mathrm{cond}_{|K\boldsymbol{b}|}((KA)(\Omega'_{QS}), \boldsymbol{x}) = \mathrm{cond}_{|\boldsymbol{b}|}(A(\Omega_{QS}), \boldsymbol{x}),$$

*where $\Omega'_{QS}$ is any quasiseparable representation of $KA$.*

*Proof.* Since $K$ is a diagonal matrix, all the following equalities are straightforward:

1. $|(KA)^{-1}|\,|K\boldsymbol{b}| = |A^{-1}|\,|K^{-1}|\,|K|\,|\boldsymbol{b}| = |A^{-1}|\,|\boldsymbol{b}|$,
2. $|(KA)^{-1}|\,|(KA_D)|\,|\boldsymbol{x}| = |A^{-1}|\,|K^{-1}|\,|K|\,|A_D|\,|\boldsymbol{x}| = |A^{-1}|\,|A_D|\,|\boldsymbol{x}|$,
3. $|(KA)^{-1}|\,|(KA_L)\boldsymbol{x}| = |A^{-1}|\,|K^{-1}|\,|K|\,|A_L\boldsymbol{x}| = |A^{-1}|\,|A_L\boldsymbol{x}|$,
4. $|(KA)^{-1}(KA_L)|\,|\boldsymbol{x}| = |A^{-1}A_L|\,|\boldsymbol{x}|$,
5. $|(KA)^{-1}|\,|(KA_U)\boldsymbol{x}| = |A^{-1}|\,|K^{-1}|\,|K|\,|A_U\boldsymbol{x}| = |A^{-1}|\,|A_U\boldsymbol{x}|$,
6. $|(KA)^{-1}(KA_U)|\,|\boldsymbol{x}| = |A^{-1}A_U|\,|\boldsymbol{x}|$,
7. $(KA)^{-1}\begin{bmatrix} 0 & 0 \\ (KA)(i+1:n, 1:i-1) & 0 \end{bmatrix}\boldsymbol{x} = A^{-1}\begin{bmatrix} 0 & 0 \\ A(i+1:n, 1:i-1) & 0 \end{bmatrix}\boldsymbol{x}$,
8. $(KA)^{-1}\begin{bmatrix} 0 & (KA)(1:j-1, j+1:n) \\ 0 & 0 \end{bmatrix}\boldsymbol{x} = A^{-1}\begin{bmatrix} 0 & A(1:j-1, j+1:n) \\ 0 & 0 \end{bmatrix}\boldsymbol{x}$.

The result follows trivially from 1)-8), $KA = KA_L + KA_D + KA_U$, Theorem 4.3, and the fact that $\mathrm{cond}_{\boldsymbol{f}}(A(\Omega_{QS}), \boldsymbol{x})$ does not depend on the particular quasiseparable parameterization used. □

## 4.2 Condition number of the solution of $\{1,1\}$-quasiseparable linear systems in the Givens-vector representation

The main goal of this section is to find an explicit expression for computing the componentwise condition number for the solution of a linear system of equations with a quasiseparable matrix of coefficients with respect to the Givens-vector representation. In Section 4.2.1, we define the Givens-vector representation via tangents and in Section 4.2.2, we provide the desired expression for evaluating the condition number with respect to that representation.

### 4.2.1 The Givens-vector representation via tangents for $\{1,1\}$-quasiseparable matrices

Since the Givens-vector representation is a particular case of a quasiseparable representation, one might think that it makes no sense to study again condition numbers for the solution of a linear system because we know, from Proposition 4.4, that they are independent of the particular choice of $\Omega_{QS}$. However, the subtle point here is that the Givens-vector representation presented in Theorem 2.21 has correlated parameters since the pairs $\{c_i, s_i\}$ are not independent; the same happens for $\{r_i, t_i\}$. Since arbitrary componentwise perturbations of $\Omega_{QS}^{GV}$ (see Theorem 2.21) destroy the cosine-sine pairs, and we want to restrict ourselves to perturbations that preserve the cosine-sine pairs, an additional parametrization of these pairs is needed.

Avoiding the use of trigonometric functions, we essentially have two options for these additional parameters:

(a) We can consider $\{c_i, s_i\} = \left\{\sqrt{1-s_i^2}, s_i\right\}_{i=2}^{n-1}$ and $\{r_i, t_i\} = \left\{\sqrt{1-t_i^2}, t_i\right\}_{i=2}^{n-1}$, but this is not convenient because if $s_i$ is too close to 1, then tiny relative variations of $s_i$ may produce large relative variations of $c_i$, since:

$$\frac{s_i}{c_i}\frac{\partial c_i}{\partial s_i} = -\frac{s_i^2}{c_i\sqrt{1-s_i^2}} = -\left(\frac{s_i}{c_i}\right)^2.$$

This is reflected in the following derivative appearing in the condition number developed in Theorem 3.10 (recall from the proof of Theorem 4.2 that we rewrite the terms $|A^{-1}\,(\partial A/\partial\omega_k)\boldsymbol{x}|\,e_k$ in the expression in Theorem 3.10 as $|A^{-1}\,\omega_k(\partial A/\partial\omega_k)\boldsymbol{x}|\,|e_k/\omega_k|$),

$$s_i\frac{\partial A}{\partial s_i} = \begin{bmatrix} 0 & 0 \\ -\left(\frac{s_i}{c_i}\right)^2 A(i,1:i-1) & 0 \\ A(i+1:n,1:i-1) & 0 \end{bmatrix}\boldsymbol{x},$$

which may be huge if $\left(\frac{s_i}{c_i}\right)^2$ is huge. The same happens for $r_i$.

(b) On the other hand, we can use tangents as parameters in the following way:

$$c_i = \frac{1}{\sqrt{1+l_i^2}}\,,\ s_i = \frac{l_i}{\sqrt{1+l_i^2}}\ ,\ \text{and } r_i = \frac{1}{\sqrt{1+u_i^2}}\,,\ t_i = \frac{u_i}{\sqrt{1+u_i^2}}\,,$$

for $i = 2,\ldots,n-1$.
Observe that when using the tangents as parameters, the value $c_i = 0$ (resp. $r_i = 0$) corresponds to $l_i = \infty$ (resp. $u_i = \infty$). In addition, recall that, since we are interested in calculating relative-relative componentwise condition

numbers, the parameters that are zero must remain zero. This is consistent with the fact that $\frac{\partial c_i}{\partial l_i}$ and $\frac{\partial s_i}{\partial l_i}$ both tend to zero when $l_i \to \infty$. The same happens with $\frac{\partial r_i}{\partial u_i}$ and $\frac{\partial t_i}{\partial u_i}$ when $u_i \to \infty$.

Since it is obvious from (b) that tiny relative perturbations of the tangent parameters $l_i$ and $u_i$ produce tiny relative perturbations of the cosine-sine parameters $\{c_i, s_i\}$ and $\{r_i, t_i\}$, respectively, it is natural to use tangent parameters in practical numerical situations. This is related to and inspired by the fact that Givens rotations are computed in practice by using tangents or cotangents. See on this point the classical reference [38, Algorithm 5.1.3] and the state-of-the-art algorithm in [8].

**Definition 4.8.** *For any Givens-vector representation*

$$\Omega_{QS}^{GV} = \left(\{c_i, s_i\}_{i=2}^{n-1}, \{v_i\}_{i=1}^{n-1}, \{d_i\}_{i=1}^{n}, \{e_i\}_{i=1}^{n-1}, \{t_i, r_i\}_{i=2}^{n-1}\right)$$

*of a* $\{1,1\}$*-quasiseparable matrix* $C \in \mathbb{R}^{n\times n}$*, we define the* Givens-vector representation via tangents *as*

$$\Omega_{GV} := \left(\{l_i\}_{i=2}^{n-1}, \{v_i\}_{i=1}^{n-1}, \{d_i\}_{i=1}^{n}, \{e_i\}_{i=1}^{n-1}, \{u_i\}_{i=2}^{n-1}\right) \in \mathbb{R}^{5n-6}, \text{ where}$$

$$c_i = \frac{1}{\sqrt{1+l_i^2}}, \; s_i = \frac{l_i}{\sqrt{1+l_i^2}}, \quad \text{and} \quad r_i = \frac{1}{\sqrt{1+u_i^2}}, \; t_i = \frac{u_i}{\sqrt{1+u_i^2}},$$

*for* $i = 2, \ldots, n-1$.

### 4.2.2 The condition number for $\{1,1\}$-quasiseparable matrices in the Givens-vector representation via tangents

In order to use differential calculus to deduce an explicit expression of the structured condition number of the solution of a linear system with respect to the tangent-Givens-vector representation, we will need Lemma 4.9.

**Lemma 4.9.** *Let* $A \in \mathbb{R}^{n\times n}$ *be a* $\{1,1\}$*-quasiseparable matrix and let* $\Omega_{GV}$ *be the tangent-Givens-vector representation of* $A$*, where* $\Omega_{GV} = \left(\{l_i\}_{i=2}^{n-1}, \{v_i\}_{i=1}^{n-1}, \{d_i\}_{i=1}^{n}, \{e_i\}_{i=1}^{n-1}, \{u_i\}_{i=2}^{n-1}\right)$*. Then the entries of* $A$ *are differentiable functions of the parameters in* $\Omega_{GV}$ *and*

*a)* $$l_i \frac{\partial A}{\partial l_i} = \begin{bmatrix} 0 & 0 \\ -s_i^2 A(i, 1:i-1) & 0 \\ c_i^2 A(i+1:n, 1:i-1) & 0 \end{bmatrix}, \text{ for } i = 2, \ldots, n-1,$$

*b)* $$u_i \frac{\partial A}{\partial u_i} = \begin{bmatrix} 0 & -t_i^2 A(1:i-1, i) & r_i^2 A(1:i-1, i+1:n) \\ 0 & 0 & 0 \end{bmatrix}, \text{ for } i = 2, \ldots, n-1.$$

*Proof.* For the derivatives with respect to the parameters $\{l_i\}_{i=2}^{n-1}$, note first that

$$l_i \frac{\partial c_i}{\partial l_i} = -\frac{l_i^2}{(1+l_i^2)^{3/2}} = -s_i^2 c_i\,, \quad l_i \frac{\partial s_i}{\partial l_i} = \frac{l_i}{(1+l_i^2)^{3/2}} = c_i^2 s_i,$$

and, therefore, from Theorem 2.21, we obtain

$$l_i \frac{\partial A}{\partial l_i} = \begin{bmatrix} 0 & 0 \\ -s_i^2 A(i, 1:i-1) & 0 \\ c_i^2 A(i+1:n, 1:i-1) & 0 \end{bmatrix}, \quad \text{for} \quad i = 2, \ldots, n-1.$$

Analogously, for the derivatives with respect to the parameters $\{u_i\}_{i=2}^{n-1}$, we have

$$u_j \frac{\partial r_j}{\partial u_j} = -\frac{u_j^2}{(1+u_j^2)^{3/2}} = -t_j^2 r_j\,, \quad u_j \frac{\partial t_j}{\partial u_j} = \frac{u_j}{(1+u_j^2)^{3/2}} = r_j^2 t_j,$$

from which we obtain that

$$u_i \frac{\partial A}{\partial u_i} = \begin{bmatrix} 0 & -t_i^2 A(1:i-1, i) & r_i^2 A(1:i-1, i+1:n) \\ 0 & 0 & 0 \end{bmatrix}, \quad \text{for} \quad i = 2, \ldots, n-1.$$

□

**Remark 4.10.** *For the partial derivatives with respect to the parameters in $\{d_i\}_{i=1}^{n}$, $\{v_i\}_{i=1}^{n-1}$, and $\{e_i\}_{i=1}^{n-1}$ see a), d), e) in Lemma 4.1. Recall that those parameters can also be respectively seen as the parameters $\{d_i\}_{i=1}^{n}$, $\{q_i\}_{i=1}^{n-1}$, and $\{g_i\}_{i=1}^{n-1}$ in a quasiseparable representation of A, as we explained in the paragraph after Theorem 2.21.*

Theorem 4.11 is the main result of Section 4.2.2 and it presents an explicit expression of the componentwise condition number for the solution of a linear system of equations with a quasiseparable matrix of coefficients with respect to the Givens-vector representation via tangents. This result is based again in Theorem 3.10.

**Theorem 4.11.** *Let $A\boldsymbol{x} = \boldsymbol{b}$, where $0 \neq \boldsymbol{x} \in \mathbb{R}^n$ and $A \in \mathbb{R}^{n\times n}$ is a nonsingular $\{1,1\}$-quasiseparable matrix with a tangent-Givens-vector representation $\Omega_{GV}$, and such that $A = A_L + A_D + A_U$, with $A_L$ strictly lower triangular, $A_D$ diagonal, and $A_U$ strictly upper triangular. Let $0 \leq \boldsymbol{f} \in \mathbb{R}^n$ and $0 \leq E_{GV} \in \mathbb{R}^{5n-6}$. Then*

$$\begin{aligned} \mathrm{cond}_{E_{GV},\boldsymbol{f}}(A(\Omega_{GV}), \boldsymbol{x}) = \frac{1}{\|\boldsymbol{x}\|_\infty} \Bigg\| & \left|A^{-1}\right| \boldsymbol{f} + |A^{-1}|Q_d|\boldsymbol{x}| + |A^{-1}A_L||Q_v||\boldsymbol{x}| \\ & + |A^{-1}||Q_e||A_U\boldsymbol{x}| + \sum_{i=2}^{n-1} \left| A^{-1} \begin{bmatrix} 0 & 0 \\ -s_i^2 A(i, 1:i-1) & 0 \\ c_i^2 A(i+1:n, 1:i-1) & 0 \end{bmatrix} \boldsymbol{x} \right| \left| \frac{e_{l_i}}{l_i} \right| \end{aligned}$$

$$+\sum_{j=2}^{n-1}\left|A^{-1}\begin{bmatrix}0 & -t_j^2A(1:j-1,j) & r_j^2A(1:j-1,j+1:n)\\ 0 & 0 & 0\end{bmatrix}\boldsymbol{x}\right|\left|\frac{e_{u_j}}{u_j}\right|\Bigg\|_\infty,$$

*where*

$$\begin{aligned}\Omega_{GV} &= \left(\{l_i\}_{i=2}^{n-1},\{v_i\}_{i=1}^{n-1},\{d_i\}_{i=1}^{n},\{e_i\}_{i=1}^{n-1},\{u_i\}_{i=2}^{n-1}\right),\ \textit{as in Definition 4.8,}\\ E_{GV} &= \left(\{e_{l_i}\}_{i=2}^{n-1},\{e_{v_i}\}_{i=1}^{n-1},\{e_{d_i}\}_{i=1}^{n},\{e_{e_i}\}_{i=1}^{n-1},\{e_{u_i}\}_{i=2}^{n-1}\right),\\ Q_d &= \text{diag}\,(e_{d_1},\ldots,e_{d_n}),\ Q_v = \text{diag}\left(\frac{e_{v_1}}{v_1},\ldots,\frac{e_{v_{n-1}}}{v_{n-1}},1\right),\\ Q_e &= \text{diag}\left(\frac{e_{e_1}}{e_1},\ldots,\frac{e_{e_{n-1}}}{e_{n-1}},1\right),\end{aligned}$$

*and each quotient whose denominator is zero must be understood as zero if the numerator is also zero and, otherwise, the zero parameter in the denominator should be formally cancelled out with the same parameter in the corresponding piece of A.*

*Proof.* The proof is straightforward from Theorem 3.10, Lemma 4.9, Remark 4.10, and the proof of Theorem 4.2. Therefore, we omit the proof. □

For the most natural choice $E_{GV} = |\Omega_{GV}|$, we adopt the shorter notation

$$\text{cond}_{\boldsymbol{f}}(A(\Omega_{GV}),\boldsymbol{x}) \equiv \text{cond}_{|\Omega_{GV}|,\boldsymbol{f}}(A(\Omega_{GV}),\boldsymbol{x}),$$

and get Theorem 4.12 as a corollary of Theorem 4.11.

**Theorem 4.12.** *Let $A\boldsymbol{x} = \boldsymbol{b}$, where $0 \neq \boldsymbol{x} \in \mathbb{R}^n$ and $A \in \mathbb{R}^{n\times n}$ is a nonsingular $\{1,1\}$-quasiseparable matrix with a tangent-Givens-vector representation $\Omega_{GV}$, and such that $A = A_L + A_D + A_U$, with $A_L$ strictly lower triangular, $A_D$ diagonal, and $A_U$ strictly upper triangular. Let $0 \leq \boldsymbol{f} \in \mathbb{R}^n$. Then*

$$\begin{aligned}&\text{cond}_{\boldsymbol{f}}(A(\Omega_{GV}),\boldsymbol{x}) =\\ &\frac{1}{\|\boldsymbol{x}\|_\infty}\Big\|\,|A^{-1}|\,\boldsymbol{f} + |A^{-1}||A_D||\boldsymbol{x}| + |A^{-1}A_L||\boldsymbol{x}| + |A^{-1}||A_U\boldsymbol{x}|\\ &+\sum_{i=2}^{n-1}\left|A^{-1}\begin{bmatrix}0 & 0\\ -s_i^2A(i,1:i-1) & 0\\ c_i^2A(i+1:n,1:i-1) & 0\end{bmatrix}\boldsymbol{x}\right|\\ &+\sum_{j=2}^{n-1}\left|A^{-1}\begin{bmatrix}0 & -t_j^2A(1:j-1,j) & r_j^2A(1:j-1,j+1:n)\\ 0 & 0 & 0\end{bmatrix}\boldsymbol{x}\right|\Bigg\|_\infty.\end{aligned}$$

Note, from the expressions in Theorems 4.12 and 4.3, that there exists an important difference between the structured condition number in the Givens-vector

representation and the structured condition number in the quasiseparable representation for a given $\{1,1\}$-quasiseparable matrix $A$, since $\mathrm{cond}_{\boldsymbol{f}}(A(\Omega_{GV}),\boldsymbol{x})$ depends not only on the entries of the matrix $A$ but on the parameters $\{c_i,s_i\}$ and $\{r_i,t_i\}$ as well, while $\mathrm{cond}_{\boldsymbol{f}}(A(\Omega_{QS}),\boldsymbol{x})$ only depends on the matrix entries. Furthermore, since the cosine-sine parameters in the Givens-vector representation do not change trivially under diagonal scalings, we have that, for the natural choice $\boldsymbol{f}=|\boldsymbol{b}|$, $\mathrm{cond}_{\boldsymbol{f}}(A(\Omega_{GV}),\boldsymbol{x})$ is not invariant under row scaling while $\mathrm{cond}_{\boldsymbol{f}}(A(\Omega_{QS}),\boldsymbol{x})$ is invariant under row scaling (recall Proposition 4.7).

## 4.3 Comparison of condition numbers in the quasiseparable and the Givens-vector representations

It is natural to expect the Givens-vector representation via tangents to be a more stable representation than the general quasiseparable representation for computing the solution of a quasiseparable linear system of equations, in the sense that the Givens-vector representation should lead to smaller condition numbers. This is due to the fact that any Givens-vector representation is also a quasiseparable representation and the perturbations considered in the condition numbers preserve the structure of the tangent-Givens-vector parametrization. In Theorem 4.13, the corresponding result is proved for linear systems, that is, that $\mathrm{cond}_{\boldsymbol{f}}(A(\Omega_{GV}),\boldsymbol{x})$ can not be larger than $\mathrm{cond}_{\boldsymbol{f}}(A(\Omega_{QS}),\boldsymbol{x})$.

**Theorem 4.13.** *Let $A\boldsymbol{x}=\boldsymbol{b}$, where $0\neq\boldsymbol{x}\in\mathbb{R}^n$ and $A\in\mathbb{R}^{n\times n}$ is a nonsingular $\{1,1\}$-quasiseparable matrix, and let $\Omega_{GV}$ be the vector of tangent-Givens-vector parameters of $A$. Then, for $0\leq\boldsymbol{f}\in\mathbb{R}^n$, and for any vector $\Omega_{QS}$ of quasiseparable parameters of $A$, the following inequality holds:*

$$\mathrm{cond}_{\boldsymbol{f}}(A(\Omega_{GV}),\boldsymbol{x})\leq\mathrm{cond}_{\boldsymbol{f}}(A(\Omega_{QS}),\boldsymbol{x}).$$

*Proof.* Throughout the proof, we use the decomposition $A=A_L+A_D+A_U$ introduced in Theorems 4.2 and 4.11. For the sums in the last two terms of the expression for $\mathrm{cond}_{\boldsymbol{f}}(A(\Omega_{GV}),\boldsymbol{x})$ we have,

$$\begin{aligned}
S_1 &:= \sum_{i=2}^{n-1}\left|A^{-1}\begin{bmatrix}0 & 0\\ -s_i^2\,A(i,1:i-1) & 0\\ c_i^2\,A(i+1:n,1:i-1) & 0\end{bmatrix}\boldsymbol{x}\right|\\
&\leq \sum_{i=2}^{n-1}\left|A^{-1}\begin{bmatrix}0 & 0\\ A(i,1:i-1) & 0\\ 0 & 0\end{bmatrix}\boldsymbol{x}\right|+\sum_{i=2}^{n-1}\left|A^{-1}\begin{bmatrix}0 & 0\\ 0 & 0\\ A(i+1:n,1:i-1) & 0\end{bmatrix}\boldsymbol{x}\right|\\
&= \sum_{i=2}^{n-1}\left|A^{-1}(:,i)\right|\left|A_L(i,:)\boldsymbol{x}\right|+\sum_{i=2}^{n-1}\left|A^{-1}\begin{bmatrix}0 & 0\\ A(i+1:n,1:i-1) & 0\end{bmatrix}\boldsymbol{x}\right|
\end{aligned}$$

$$\leq \left|A^{-1}\right| |A_L \boldsymbol{x}| + \sum_{i=2}^{n-1} \left| A^{-1} \begin{bmatrix} 0 & 0 \\ A(i+1:n, 1:i-1) & 0 \end{bmatrix} \boldsymbol{x} \right| \tag{4.1}$$

and, in an analogous way,

$$\begin{aligned} S_2 &:= \sum_{j=2}^{n-1} \left| A^{-1} \begin{bmatrix} 0 & -t_j^2\, A(1:j-1, j) & r_j^2\, A(1:j-1, j+1:n) \\ 0 & 0 & 0 \end{bmatrix} \boldsymbol{x} \right| \\ &\leq \left|A^{-1} A_U\right| |\boldsymbol{x}| + \sum_{j=2}^{n-1} \left| A^{-1} \begin{bmatrix} 0 & A(1:j-1, j+1:n) \\ 0 & 0 \end{bmatrix} \boldsymbol{x} \right|. \end{aligned} \tag{4.2}$$

From (4.1) and (4.2) the proof is straightforward by comparing the expressions in Theorem 4.3 and Theorem 4.12 for $\mathrm{cond}_{\boldsymbol{f}}(A(\Omega_{QS}), \boldsymbol{x})$ and $\mathrm{cond}_{\boldsymbol{f}}(A(\Omega_{GV}), \boldsymbol{x})$, respectively. □

On the other hand, as we prove in Theorem 4.16 below, the Givens-vector representation via tangents can only improve the relative condition number for the solution of a $\{1,1\}$-quasiseparable linear system of equations up to a factor of $3n$ with respect to the quasiseparable representation. Therefore, we conclude that, when computing solutions of $\{1,1\}$-quasiseparable linear systems of equations, these representations can be considered numerically equivalent in terms of expected accuracy.

For proving Theorem 4.16, we need the simple Lemma 4.14. It is worth to observe that results in the spirit of Lemma 4.14 can be found in the detailed error analysis of Givens rotations presented in [4].

**Lemma 4.14.** *Let $l \neq 0$ be a real number representing a tangent and $c$ be the corresponding positive cosine. Then, for any positive value $\eta < 1$, a relative perturbation of $l$ by at most $\eta$ produces a relative perturbation of $c$ of the order of $\eta$, i.e.,*

$$\left|\frac{\delta l}{l}\right| \leq \eta \implies \left|\frac{\delta c}{c}\right| \leq \left(\eta + \mathcal{O}(\eta^2)\right), \text{ where } c + \delta c = \frac{1}{\sqrt{1 + (l + \delta l)^2}}.$$

*Proof.* Consider $l' = l + \delta l$ as the perturbed tangent and $c' = c + \delta c = 1/\sqrt{1 + (l')^2}$ as the respective perturbed cosine . Then, for $1 > \eta > 0$ sufficiently small we have that if $(1-\eta)|l| \leq |l'| \leq (1+\eta)|l|$, then

$$\frac{c}{1+\eta} = \frac{1}{(1+\eta)(\sqrt{1+l^2})} \leq c' \leq \frac{1}{(1-\eta)(\sqrt{1+l^2})} = \frac{c}{1-\eta},$$

from where we can conclude that $|\delta l| \leq \eta |l| \Rightarrow |\delta c| \leq (\eta + \mathcal{O}(\eta^2))|c|$. □

In order to prove Theorem 4.16, we develop a proof based on Definition 3.7. Recall that from the Givens-vector representation via tangents $\Omega_{GV}$ of $A$ we can obtain the Givens-vector representation $\Omega_{QS}^{GV}$ of $A$ as in Definition 4.8, and that

$\Omega_{QS}^{GV}$ is also a quasiseparable representation of $A$. Therefore, in order to use the componentwise relative condition numbers for representations in Definition 3.7, let us consider a quasiseparable perturbation $\delta\Omega_{QS}^{GV}$ of the parameters in $\Omega_{QS}^{GV}$ such that $|\delta\Omega_{QS}^{GV}| \leq \eta|\Omega_{QS}^{GV}|$, and the resulting quasiseparable matrix $\tilde{A} = A(\Omega_{QS}^{GV} + \delta\Omega_{QS}^{GV})$ (note that the perturbations $\delta\Omega_{QS}^{GV}$ do not respect in general the pairs cosine-sine of $\Omega_{QS}^{GV}$). We will refer to $\eta$ as the *level* of the relative perturbation of the parameters in the representation $\Omega_{QS}^{GV}$. Moreover, note that $\tilde{A}$ can also be represented by a vector

$$\Omega'_{GV} := \left(\{l'_i\}_{i=2}^{n-1}, \{v'_i\}_{i=1}^{n-1}, \{d'_i\}_{i=1}^{n}, \{e'_i\}_{i=1}^{n-1}, \{u'_i\}_{i=2}^{n-1}\right)$$

of tangent-Givens-vector parameters and let us consider the perturbations $\delta'\Omega_{GV} := \Omega'_{GV} - \Omega_{GV}$. Then, Lemma 4.15 states a bound for the level $\eta'$ of the respective relative perturbation over the parameters in $\Omega'_{GV}$ produced by a relative perturbation of level $\eta$ over the quasiseparable parameters in $\Omega_{QS}^{GV}$.

**Lemma 4.15.** *Let $A$ be a $\{1,1\}$-quasiseparable matrix with Givens-vector representation via tangents $\Omega_{GV}$. Then, using the notation in the previous paragraph, and for $\eta$ sufficiently small, we have:*

$$|\delta\Omega_{QS}^{GV}| \leq \eta|\Omega_{QS}^{GV}| \Longrightarrow |\delta'\Omega_{GV}| \leq (3(n-2)\eta + \mathcal{O}(\eta^2))|\Omega_{GV}|.$$

*Proof.* Let us proceed by analyzing each subset of parameters as follows.

- For the parameters in $\{d'_i\}_{i=1}^{n}$ it is obvious that $\delta' d_i = \delta d_i$.
- For the parameters in $\{v'_j\}_{j=1}^{n-1}$ note that $v'_j = \sqrt{\sum_{i=j+1}^{n} (\tilde{A}(i,j))^2}$ from where it is easy to see that if $|\delta\Omega_{QS}^{GV}| \leq \eta|\Omega_{QS}^{GV}|$, then

$$\sqrt{\sum_{i=j+1}^{n} [(1-\eta)^n A(i,j)]^2} \leq v'_j \leq \sqrt{\sum_{i=j+1}^{n} [(1+\eta)^n A(i,j)]^2},$$

$$(1-\eta)^n \sqrt{\sum_{i=j+1}^{n} [A(i,j)]^2} \leq v'_j \leq (1+\eta)^n \sqrt{\sum_{i=j+1}^{n} [A(i,j)]^2},$$

$$(1-\eta)^n v_j \leq v'_j \leq (1+\eta)^n v_j,$$

  and, consequently, we have that $|\delta' v_j| \leq (n\eta + \mathcal{O}(\eta^2))|v_j|$, for $j = 1, \ldots, n-1$.
- For studying the level of perturbation in the parameters $\{l'_i\}_{i=2}^{n-1}$ produced by the level $\eta$ of perturbation in $\Omega_{QS}^{GV} + \delta\Omega_{QS}^{GV}$, recall first that the parameter $p_n = 1$ of $\Omega_{QS}^{GV}$ is also perturbed. Next, taking into account which are the entries of $\tilde{A}$

in the representations $\Omega_{QS}^{GV} + \delta\Omega_{QS}^{GV}$ and $\Omega'_{GV}$ according to Theorems 2.18 and 2.21, observe

$$\begin{aligned} l'_{n-1} &= \frac{s'_{n-1}}{c'_{n-1}} = \frac{\tilde{C}(n, n-2)}{\tilde{C}(n-1, n-2)} = \frac{(1+\epsilon_{n-1})s_{n-1}(1+\alpha_{n-1})}{c_{n-1}(1+\beta_{n-1})}, \quad \text{and} \\ l'_i &= \frac{s'_i}{c'_i} = \frac{\tilde{C}(i+1, i-1)}{\tilde{C}(i, i-1)} \frac{1}{c'_{i+1}} = \frac{c_{i+1}(1+\epsilon_i)s_i(1+\alpha_i)}{c_i(1+\beta_i)c'_{i+1}}, \end{aligned} \tag{4.3}$$

for $i = n-2, n-3, \ldots, 2$, where

$$|\epsilon_i| \leq \eta,\ |\alpha_i| \leq \eta,\ |\beta_i| \leq \eta, \ \text{ for } \ i = n-1, n-2, \ldots, 2.$$

From Lemma 4.14, we know that if $|\delta' l_{i+1}| = |l'_{i+1} - l_{i+1}| \leq \kappa_{i+1}|l_{i+1}|$, then $|\delta' c_{i+1}| = |c'_{i+1} - c_{i+1}| \leq \left(\kappa_{i+1} + \mathcal{O}(\kappa_{i+1}^2)\right)|c_{i+1}|$, or equivalently $c'_{i+1} = c_{i+1}(1+\theta_{i+1})$, with $|\theta_{i+1}| \leq \kappa_{i+1} + \mathcal{O}(\kappa_{i+1}^2)$. Therefore, the second equation in (4.3) implies

$$|\delta' l_i| = |l'_i - l_i| \leq (3\eta + \kappa_{i+1} + \mathcal{O}(\eta^2 + \kappa_{i+1}^2))|l_i| \ \text{ for } \ i = n-2, n-3, \ldots, 2, \tag{4.4}$$

and the first equation in (4.3) implies $|\delta' l_{n-1}| \leq (3\eta + \mathcal{O}(\eta^2))|l_{n-1}|$, i.e., up to first order in $\eta$, we can take $\kappa_{n-1} = 3\eta$. So, equation (4.4) provides a recurrence relation for $\kappa_{i+1}$ and we get

$$|\delta' l_i| \leq (3(n-i)\eta + \mathcal{O}(\eta^2))|l_i|.$$

- For the parameters in $\{e'_i\}_{i=1}^{n-1}$, we can proceed in an analogous way to that for the parameters in $\{v'_j\}_{j=1}^{n-1}$, and we obtain that $|\delta' e_i| \leq (n\eta + \mathcal{O}(\eta^2))|e_i|$, for $i = 1, \ldots, n-1$.

- For the parameters in $\{u'_j\}_{j=2}^{n-1}$, we can also proceed in an analogous way to that for the parameters in $\{l'_i\}_{i=2}^{n-1}$, and we obtain that $|\delta' u_j| \leq (3(n-j)\eta + \mathcal{O}(\eta^2))|u_j|$, for $j = 2, \ldots, n-1$.

The proof is completed by pointing out that the desired result follows trivially from the bounds obtained above for the levels of relative perturbations with respect to the different sets of parameters in $\Omega'_{GV}$. □

**Theorem 4.16.** *Let $A\boldsymbol{x} = \boldsymbol{b}$, where $0 \neq \boldsymbol{x} \in \mathbb{R}^n$ and $A \in \mathbb{R}^{n\times n}$ is a nonsingular $\{1,1\}$-quasiseparable matrix with tangent-Givens-vector representation $\Omega_{GV}$. Let $0 \leq \boldsymbol{f} \in \mathbb{R}^n$. Then, for any quasiseparable representation $\Omega_{QS}$ of $A$:*

$$\frac{\text{cond}_{\boldsymbol{f}}(A(\Omega_{QS}), \boldsymbol{x})}{\text{cond}_{\boldsymbol{f}}(A(\Omega_{GV}), \boldsymbol{x})} \leq 3(n-2).$$

*Proof.* Note that from Definition 3.7 and from Lemma 4.15 we have

$$\begin{aligned}\mathrm{cond}_{\boldsymbol{f}}(A(\Omega_{QS}),\boldsymbol{x}) \leq \lim_{\eta\to 0}\sup\Big\{ \frac{\|\delta\boldsymbol{x}\|_\infty}{\eta\|\boldsymbol{x}\|_\infty} : &(A(\Omega_{GV}+\delta\Omega_{GV}))\,(\boldsymbol{x}+\delta\boldsymbol{x}) = \boldsymbol{b}+\delta\boldsymbol{b},\\ &|\delta\Omega_{GV}| \leq (3(n-2)\eta+\mathcal{O}(\eta^2))|\Omega_{GV}|,\\ &|\delta\boldsymbol{b}| \leq (3(n-2)\eta+\mathcal{O}(\eta^2))\boldsymbol{f}\Big\}.\end{aligned}$$

By considering the change of variable $\eta' = (3(n-2)\eta+\mathcal{O}(\eta^2))$, we obtain

$$\begin{aligned}\mathrm{cond}_{\boldsymbol{f}}(A(\Omega_{QS}),\boldsymbol{x}) &\leq \lim_{\eta'\to 0}\sup\left\{ \frac{3(n-2)\|\delta\boldsymbol{x}\|_\infty}{\eta'\|\boldsymbol{x}\|_\infty} : (A(\Omega_{GV}+\delta\Omega_{GV}))\,(\boldsymbol{x}+\delta\boldsymbol{x}) = \boldsymbol{b}+\delta\boldsymbol{b},\right.\\ &\qquad\left. |\delta\Omega_{GV}| \leq \eta'|\Omega_{GV}|, |\delta\boldsymbol{b}| \leq \eta'\boldsymbol{f}\right\}\\ &= 3(n-2)\mathrm{cond}_{\boldsymbol{f}}(A(\Omega_{GV}),\boldsymbol{x}).\end{aligned}$$

□

## 4.4 Fast estimation of condition numbers: the effective condition number

To compute $\mathrm{cond}_{\boldsymbol{f}}(A(\Omega_{QS}),\boldsymbol{x})$ and $\mathrm{cond}_{\boldsymbol{f}}(A(\Omega_{GV}),\boldsymbol{x})$ fast, i.e., in $\mathcal{O}(n)$ flops, is not easy because of the two sums that appear in the expressions in Theorems 4.3 and 4.12, respectively, since they require to compute a sum of $n$ vectors, which cost $\mathcal{O}(n^2)$ flops, in addition to other computations. Then, a natural question now is to determine whether or not the contributions of these sums to the condition numbers in which they arise are significant. This question is answered in Theorems 4.18 and 4.19, in which we provide upper and lower bounds for $\mathrm{cond}_{\boldsymbol{f}}(A(\Omega_{QS}),\boldsymbol{x})$ and $\mathrm{cond}_{\boldsymbol{f}}(A(\Omega_{GV}),\boldsymbol{x})$ respectively, in terms of the *effective condition number* in Definition 4.17. We will show in this way that such effective condition number contains the essential terms in the expressions given in Theorems 4.3 and 4.12.

**Definition 4.17.** *Let $A\boldsymbol{x} = \boldsymbol{b}$, where $0 \neq \boldsymbol{x} \in \mathbb{R}^n$ and $A \in \mathbb{R}^{n\times n}$ is a nonsingular $\{1,1\}$-quasiseparable matrix such that $A = A_L + A_D + A_U$, with $A_L$ strictly lower triangular, $A_D$ diagonal, and $A_U$ strictly upper triangular. Let $0 \leq \boldsymbol{f} \in \mathbb{R}^n$. Then, for any quasiseparable representation $\Omega_{QS}$ of $A$, we define the* effective relative condition number $\mathrm{condeff}_{\boldsymbol{f}}(A(\Omega_{QS}),\boldsymbol{x})$ *for the solution of $A\boldsymbol{x} = \boldsymbol{b}$ as*

$$\begin{aligned}\mathrm{condeff}_{\boldsymbol{f}}(A(\Omega_{QS}),\boldsymbol{x}) := \frac{1}{\|\boldsymbol{x}\|_\infty}\Big\| &\left|A^{-1}\right|\boldsymbol{f} + |A^{-1}||A_D||\boldsymbol{x}| + |A^{-1}||A_L\boldsymbol{x}|\\ &+ |A^{-1}A_L||\boldsymbol{x}| + |A^{-1}||A_U\boldsymbol{x}| + |A^{-1}A_U||\boldsymbol{x}|\Big\|_\infty.\end{aligned}$$

Recall from Proposition 4.4 that the condition number $\mathrm{cond}_{\boldsymbol{f}}(A(\Omega_{QS}), \boldsymbol{x})$ does not depend on the choice of a quasiseparable representation $\Omega_{QS}$, and note from Definition 4.17 that the same holds for $\mathrm{condeff}_{\boldsymbol{f}}(A(\Omega_{QS}), \boldsymbol{x})$. Therefore, this effective condition number is always the same for any vector of quasiseparable parameters representing the matrix.

**Theorem 4.18.** *Let $A\boldsymbol{x} = \boldsymbol{b}$, where $0 \neq \boldsymbol{x} \in \mathbb{R}^n$ and $A \in \mathbb{R}^{n\times n}$ is a nonsingular $\{1,1\}$-quasiseparable matrix. Let $0 \leq \boldsymbol{f} \in \mathbb{R}^n$. Then, for any quasiseparable representation $\Omega_{QS}$ of $A$, the following relations hold:*

$$\mathrm{condeff}_{\boldsymbol{f}}(A(\Omega_{QS}), \boldsymbol{x}) \leq \mathrm{cond}_{\boldsymbol{f}}(A(\Omega_{QS}), \boldsymbol{x}) \leq (n-1)\mathrm{condeff}_{\boldsymbol{f}}(A(\Omega_{QS}), \boldsymbol{x}).$$

*Proof.* Throughout the proof, we use the decomposition $A = A_L + A_D + A_U$ introduced in Definition 4.17. Note first that the left hand side inequality is trivial from the respective expressions of $\mathrm{condeff}_{\boldsymbol{f}}(A(\Omega_{QS}), \boldsymbol{x})$ in Definition 4.17 and $\mathrm{cond}_{\boldsymbol{f}}(A(\Omega_{QS}), \boldsymbol{x})$ in Theorem 4.3. On the other hand, note that

$$\begin{aligned}
&\left|A^{-1}\begin{bmatrix} 0 & 0 \\ A(i+1:n,1:i-1) & 0 \end{bmatrix}\boldsymbol{x}\right| = \left|A^{-1}\begin{bmatrix} 0 & 0 \\ A_L(i+1:n,1:i-1) & 0 \end{bmatrix}\boldsymbol{x}\right| \\
&= \left|A^{-1}\begin{bmatrix} 0 & 0 \\ A_L(i+1:n,1:i-1) & A_L(i+1:n,i:n) \end{bmatrix}\boldsymbol{x}\right. \\
&\quad \left. + A^{-1}\begin{bmatrix} 0 & 0 \\ 0 & -A_L(i+1:n,i:n) \end{bmatrix}\boldsymbol{x}\right| \\
&\leq \left|A^{-1}\right|\left|\begin{bmatrix} 0 \\ A_L(i+1:n,:) \end{bmatrix}\boldsymbol{x}\right| \\
&\quad + \left|A^{-1}\begin{bmatrix} 0 & 0 \\ 0 & A_L(i+1:n,i:n) \end{bmatrix}\right||\boldsymbol{x}| \\
&\leq \left|A^{-1}\right||A_L\boldsymbol{x}| + \left|A^{-1}A_L\right||\boldsymbol{x}|,
\end{aligned}$$

from where we obtain:

$$\sum_{i=2}^{n-1}\left|A^{-1}\begin{bmatrix} 0 & 0 \\ A(i+1:n,1:i-1) & 0 \end{bmatrix}\boldsymbol{x}\right| \leq (n-2)\left|A^{-1}\right||A_L\boldsymbol{x}| + (n-2)\left|A^{-1}A_L\right||\boldsymbol{x}|. \tag{4.5}$$

In an analogous way, it can be proved that

$$\sum_{j=2}^{n-1}\left|A^{-1}\begin{bmatrix} 0 & A(1:j-1,j+1:n) \\ 0 & 0 \end{bmatrix}\boldsymbol{x}\right| \leq (n-2)\left|A^{-1}\right||A_U\boldsymbol{x}| + (n-2)\left|A^{-1}A_U\right||\boldsymbol{x}|. \tag{4.6}$$

Finally, note that from the inequalities in (4.5) and (4.6), and from Theorem 4.3, it is straightforward that

$$\mathrm{cond}_{\boldsymbol{f}}(A(\Omega_{QS}), \boldsymbol{x}) \leq (n-1)\left\|\left|A^{-1}\right|\boldsymbol{f} + |A^{-1}||A_D||\boldsymbol{x}| + |A^{-1}||A_L\boldsymbol{x}|\right.$$

$$+ |A^{-1}A_L||\boldsymbol{x}| + |A^{-1}||A_U\boldsymbol{x}| + |A^{-1}A_U||\boldsymbol{x}| \Big\|_\infty / \|\boldsymbol{x}\|_\infty$$
$$= (n-1)\,\mathrm{condeff}_{\boldsymbol{f}}(A(\Omega_{QS}), \boldsymbol{x}).$$

□

**Theorem 4.19.** *Let $A\boldsymbol{x} = \boldsymbol{b}$, where $0 \neq \boldsymbol{x} \in \mathbb{R}^n$ and $A \in \mathbb{R}^{n\times n}$ is a nonsingular $\{1,1\}$-quasiseparable matrix with tangent-Givens-vector representation $\Omega_{GV}$. Let $0 \leq \boldsymbol{f} \in \mathbb{R}^n$. Then, for any quasiseparable representation $\Omega_{QS}$ of $A$, the following relations hold:*

$$\frac{\mathrm{condeff}_{\boldsymbol{f}}(A(\Omega_{QS}), \boldsymbol{x})}{3(n-2)} \leq \mathrm{cond}_{\boldsymbol{f}}(A(\Omega_{GV}), \boldsymbol{x}) \leq (n-1)\mathrm{condeff}_{\boldsymbol{f}}(A(\Omega_{QS}), \boldsymbol{x}).$$

*Proof.* It follows trivially from Theorems 4.13, 4.16 and 4.18. □

Note that from Theorems 4.18 and 4.19 we can conclude that the structured condition numbers $\mathrm{cond}_{\boldsymbol{f}}(A(\Omega_{QS}), \boldsymbol{x})$ and $\mathrm{cond}_{\boldsymbol{f}}(A(\Omega_{GV}), \boldsymbol{x})$ can be both estimated, "up to a factor of order $n$", by computing the easier expression in Definition 4.17 for the effective condition number. Next, we prove in Proposition 4.20 that this effective condition number can be computed fast by using, for instance, some previous results from [24] and [25].

**Proposition 4.20.** *Let $A\boldsymbol{x} = \boldsymbol{b}$, where $0 \neq \boldsymbol{x} \in \mathbb{R}^n$ and $A \in \mathbb{R}^{n\times n}$ is a nonsingular $\{1,1\}$-quasiseparable matrix with a quasiseparable representation $\Omega_{QS}$ which is assumed to be known. Let $0 \leq \boldsymbol{f} \in \mathbb{R}^n$. Then, the effective condition number $\mathrm{condeff}_{\boldsymbol{f}}(A(\Omega_{QS}), \boldsymbol{x})$ can be computed in $\mathcal{O}(n)$ operations, i.e., with linear complexity.*

*Proof.* This assertion is a consequence of the results in [24] and [25]. Using [25, Algorithm 5.1] we can obtain in $\mathcal{O}(n)$ flops the quasiseparable representation for the inverse of the $\{1,1\}$-quasiseparable matrix A which is also $\{1,1\}$-quasiseparable ([24, Theorem 5.2]). In addition, in [24, Algorithm 4.4] it is shown how to compute the matrix-vector product $\boldsymbol{y} = R\boldsymbol{z}$ for the general case when $R$ is an $\{n_L, n_U\}$-quasiseparable matrix with a given quasiseparable representation, with linear complexity in the size $n$ of the vector. Therefore, we can use this algorithm (twice when necessary) for computing each of the products $|A^{-1}|\,\boldsymbol{f}$, $|A^{-1}|\,(|A_D|\,|\boldsymbol{x}|)$, $|A^{-1}|\,(|A_L\boldsymbol{x}|)$, and $|A^{-1}|\,(|A_U\boldsymbol{x}|)$ (in practice, the contribution of these terms to $\mathrm{condeff}_{\boldsymbol{f}}(A(\Omega_{QS}), \boldsymbol{x})$ is computed as $|A^{-1}|\,(\boldsymbol{f} + |A_D|\,|\boldsymbol{x}| + |A_L\boldsymbol{x}| + |A_U\boldsymbol{x}|)$). On the other hand in [24, Theorem 4.1] it is proved that the product $R_1R_2$ of an $\{n_1, m_1\}$-quasiseparable matrix $R_1$ times an $\{n_2, m_2\}$-quasiseparable matrix $R_2$ is, in general, an $\{n_1+n_2, m_1+m_2\}$-quasiseparable matrix, and in [24, Algorithm 4.3] it is shown how to compute with linear complexity a quasiseparable representation for this product given the representations of the factors. Therefore, since our matrix $A$ is a $\{1,1\}$-quasiseparable

matrix, we have that the products $A^{-1}A_L$ and $A^{-1}A_U$ are both quasiseparable matrices of orders $\{2,1\}$ and $\{1,2\}$, respectively, at most, and we can compute via [24, Algorithm 4.3] their quasiseparable representations. Then, once we have obtained such representations, we can use again [24, Algoritm 4.4] for calculating the products $|A^{-1}A_L|\,|\boldsymbol{x}|$ and $|A^{-1}A_U|\,|\boldsymbol{x}|$ also with a linear cost. Then, from Definition 4.17, it is straightforward that $\mathrm{condeff}_{\boldsymbol{f}}(A(\Omega_{QS}),\boldsymbol{x})$ can be computed with linear complexity. □

Finally, note that from Definition 4.17 and from the proof of Proposition 4.7, it is straightforward to prove that the effective condition number is invariant under row scaling for $\boldsymbol{f} = |\boldsymbol{b}|$, as $\mathrm{cond}_{|\boldsymbol{b}|}(A(\Omega_{QS}),\boldsymbol{x})$. This is stated without proof in Proposition 4.21.

**Proposition 4.21.** *Let $A\boldsymbol{x} = \boldsymbol{b}$, where $0 \neq \boldsymbol{x} \in \mathbb{R}^n$ and $A \in \mathbb{R}^{n\times n}$ is a $\{1,1\}$-quasiseparable matrix, and let $K \in \mathbb{R}^{n\times n}$ be an invertible diagonal matrix. Then*

$$\mathrm{condeff}_{|\boldsymbol{b}|}(A(\Omega_{QS}),\boldsymbol{x}) = \mathrm{condeff}_{|\boldsymbol{Kb}|}((KA)(\Omega'_{QS}),\boldsymbol{x}),$$

*where $\Omega_{QS}$ is any quasiseparable representation of $A$ and $\Omega'_{QS}$ is any quasiseparable representation of $KA$.*

## 4.5 Numerical experiments

This section is devoted to describe briefly some numerical experiments that have been performed in order to complete our comparison between the structured effective condition number $\mathrm{condeff}_{|\boldsymbol{b}|}(A(\Omega_{QS}),\boldsymbol{x})$ in Definition 4.17 and the unstructured one $\mathrm{cond}_{|A|,|\boldsymbol{b}|}(A,\boldsymbol{x})$ in Definition 3.4. We have used MATLAB for running several random numerical tests. First, the command `randn` from MATLAB has been used for generating the random parameters in a quasiseparable representation for a $\{1,1\}$-quasiseparable matrix of size $n \times n$, i.e., the following random vectors of parameters are generated:

$$\boldsymbol{p}_\Omega \in \mathbb{R}^{n-1}, \boldsymbol{a}_\Omega \in \mathbb{R}^{n-2}, \boldsymbol{q}_\Omega \in \mathbb{R}^{n-1}, \boldsymbol{d}_\Omega \in \mathbb{R}^{n}, \boldsymbol{g}_\Omega \in \mathbb{R}^{n-1}, \boldsymbol{b}_\Omega \in \mathbb{R}^{n-2}, \text{ and } \boldsymbol{h}_\Omega \in \mathbb{R}^{n-1}. \tag{4.7}$$

We also generate the random right-hand side vector $\boldsymbol{b} \in \mathbb{R}^n$ by using the command `randn`. Then, we construct the matrix $A$ described by the parameters in (4.7) by using the expression in Theorem 2.18, obtain the vector of solutions $\boldsymbol{x}$ using the command $A\backslash\boldsymbol{b}$ from MATLAB, and compute the structured effective condition number $\mathrm{condeff}_{|\boldsymbol{b}|}(A(\Omega_{QS}),\boldsymbol{x})$ and the unstructured condition number $\mathrm{cond}_{|A|,|\boldsymbol{b}|}(A,\boldsymbol{x})$ by using direct matrix-vector multiplication and the `inv` command of MATLAB.

In general, when using just random parameters, we have obtained similar, often moderate, values for the effective condition number and the unstructured condition number for the solution of linear systems, i.e., $\mathrm{condeff}_{|\boldsymbol{b}|}(A(\Omega_{QS}),\boldsymbol{x}) \approx \mathrm{cond}_{|A|,|\boldsymbol{b}|}(A,\boldsymbol{x})$.

Therefore, we have performed several tests using different kinds of scalings over the generated quasiseparable parameters in order to obtain unbalanced quasiseparable matrices which may be very ill conditioned.

In particular, after generating the vectors of parameters in (4.7) and the vector $\boldsymbol{b}$, we have modified $\boldsymbol{p}_\Omega$ and $\boldsymbol{h}_\Omega$ as follows

$$\boldsymbol{p}_\Omega = k_1 * \boldsymbol{p}_\Omega \quad \text{and} \quad \boldsymbol{h}_\Omega = {k_2}^{-1} * \boldsymbol{h}_\Omega,$$

where $k_1$ and $k_2$ are fixed natural numbers not greater than $10^3$. This scaling, combined with the randomness of $\boldsymbol{p}_\Omega$ and $\boldsymbol{h}_\Omega$ and the rest of parameters, produces sometimes matrices with unbalanced lower left and upper right corners (see the matrix at the end of this section for an example), for which the unstructured condition number $\text{cond}_{|A|,|\boldsymbol{b}|}(A, \boldsymbol{x})$ can be much larger than the structured one $\text{condeff}_{|\boldsymbol{b}|}(A(\Omega_{QS}), \boldsymbol{x})$. In fact, for $n = 5$, $n = 10$, $n = 50$, and $n = 100$, we have obtained vectors of parameters generating particular matrices $A$ and vectors $\boldsymbol{b}$ such that:

| $n$ | $\frac{\text{cond}_{\lvert A\rvert,\lvert\boldsymbol{b}\rvert}(A,\boldsymbol{x})}{\text{condeff}_{\lvert\boldsymbol{b}\rvert}(A(\Omega_{QS}),\boldsymbol{x})}$ | $\text{cond}_{\lvert A\rvert,\lvert\boldsymbol{b}\rvert}(A, \boldsymbol{x})$ | $\text{condeff}_{\lvert\boldsymbol{b}\rvert}(A(\Omega_{QS}), \boldsymbol{x})$ |
|---|---|---|---|
| 5 | $1.6139 \cdot 10^4$ | $6.5318 \cdot 10^4$ | 4.0471 |
| 10 | $1.9980 \cdot 10^6$ | $3.1788 \cdot 10^7$ | 15.8768 |
| 50 | $1.4107 \cdot 10^7$ | $8.7762 \cdot 10^8$ | 62.2104 |
| 100 | $1.6804 \cdot 10^9$ | $6.0297 \cdot 10^{10}$ | 35.8823 |

,

where $\boldsymbol{x} = A \backslash \boldsymbol{b}$ in each case.

From these numerical experiments we conclude that the structured effective condition number $\text{condeff}_{|\boldsymbol{b}|}(A(\Omega_{QS}), \boldsymbol{x})$ may be indeed much smaller than the unstructured one $\text{cond}_{|A|,|\boldsymbol{b}|}(A, \boldsymbol{x})$, in other words, that there exist linear systems of equations with $\{1, 1\}$-quasiseparable matrices of coefficients that have solutions which are very ill conditioned with respect to perturbations of the entries of the matrix, but that are very well conditioned with respect to perturbations on the quasiseparable parameters representing the matrix. The particular *structure* observed in the matrices that produced such huge differences between the structured and the unstructured condition numbers is illustrated in the following matrix and the respective vector $\boldsymbol{b}$, which are the ones that produced the results in the table above for $n = 5$:

$$A = \begin{bmatrix} -7.8876 \cdot 10^{-2} & -1.3485 \cdot 10^{-2} & -7.8066 \cdot 10^{-3} & 2.7951 \cdot 10^{-3} & 5.1089 \cdot 10^{-5} \\ 3.0423 \cdot 10^{-1} & -5.6399 \cdot 10^{-1} & 1.6206 \cdot 10^{-1} & -5.8026 \cdot 10^{-2} & -1.0606 \cdot 10^{-3} \\ 5.5451 \cdot 10^{1} & -2.5873 \cdot 10^{-1} & 1.3525 \cdot 10^{0} & -1.5088 \cdot 10^{-3} & -2.7578 \cdot 10^{-5} \\ -3.8947 \cdot 10^{5} & 1.8172 \cdot 10^{3} & -1.7047 \cdot 10^{0} & 3.0944 \cdot 10^{-3} & -6.7069 \cdot 10^{-3} \\ 1.7714 \cdot 10^{8} & -8.2653 \cdot 10^{5} & 7.7535 \cdot 10^{2} & 8.3875 \cdot 10^{-1} & -2.0998 \cdot 10^{0} \end{bmatrix},$$

$$\boldsymbol{b} = \begin{bmatrix} -8.8528 \cdot 10^{-1} & -1.3154 \cdot 10^{-1} & -1.5711 \cdot 10^{0} & -7.8284 \cdot 10^{-1} & -1.0898 \cdot 10^{0} \end{bmatrix}^T.$$

Note that there is an obvious unbalance in the matrix entries, since the absolute values of the entries near the lower left corner are large compared with the absolute

values of the entries in the opposite upper right corner of the matrix (compare the absolute values of the entries of the submatrix $A(4:5, 1:2)$ versus those from $A(1:2, 4:5)$).

# Chapter 5

# Structured eigenvalue condition numbers for parameterized quasiseparable matrices

In this chapter we study eigenvalue condition numbers for $\{1,1\}$-quasiseparable matrices with respect to perturbations on the parameters in the quasiseparable representation and in the Givens vector representation. In the case of the quasiseparable representation, in Sections 5.1 and 5.2 we present the original results that we have obtained for these eigenvalue condition numbers, and they include the procedure and the main techniques that will be used through the rest of the chapter in order to obtain analogous results for the condition number with respect to the Givens-vector representation, which is covered in Sections 5.3 and 5.4. More precisely, in Theorems 5.1 and 5.11, we obtain the respective expressions for the eigenvalue condition numbers for the quasiseparable and the Givens-vector representations. In Proposition 5.4 we prove that for $\{1,1\}$-quasiseparable matrices, the structure plays an important role in the sensitivity of eigenvalues, since the quasiseparable representation yields an structured condition number which can not be much larger than the standard unstructured one, but that can be much smaller, as it is observed in the numerical experiments described in Section 5.6. In Section 5.5 we compare the eigenvalue condition numbers in the quasiseparable and in the Givens-vector representation and prove that both representations can be considered numerically equivalent in terms of the accuracy of eigenvalue computations as it is deduced from Theorems 5.15 and 5.16. That is, we obtain a result similar to the one obtained for the solutions of linear systems.

## 5.1 Eigenvalue condition numbers for $\{1,1\}$-quasiseparable matrices in the quasiseparable representation

In this section we will deduce an explicit expression for the eigenvalue condition number for $\{1,1\}$-quasiseparable matrices in the quasiseparable representation. This representation was introduced in [24] together with the definition of quasiseparable matrices, which are described in the previous Sections 2.2 and 2.3.1 of this dissertation.

From Example 2.19 it seems natural to consider relative componentwise perturbations of the vector of parameters $\Omega_{QS}$ introduced in Theorem 2.18 instead of normwise perturbations of $\Omega_{QS}$, because the norm of the vector of parameters $\Omega_{QS}$ does not determine the norm of the matrix. Since any $\{1,1\}$-quasiseparable matrix is differentiable with respect to these parameters, we can deduce eigenvalue relative-relative componentwise condition numbers for this parametrization by using Theorem 3.23. This is done in Theorem 5.1 for relative perturbations in the quasiseparable parameters $\Omega_{QS}$ with respect to a general vector of nonnegative entries.

**Theorem 5.1.** *Let $C \in \mathbb{R}^{n\times n}$ be a $\{1,1\}$-quasiseparable matrix and let us express $C$ as $C = C_L + C_D + C_U$, with $C_L$ strictly lower triangular, $C_D$ diagonal, and $C_U$ strictly upper triangular. Suppose $\lambda \neq 0$ is a simple eigenvalue of $C$ with left and right eigenvectors $\boldsymbol{y}$ and $\boldsymbol{x}$, respectively. Denote by $\Omega_{QS}$ a quasiseparable representation of $C$ as in Theorem 2.18 and let $0 \leq E_{QS} \in \mathbb{R}^{7n-8}$. Then,*

$$\begin{aligned}\operatorname{cond}_{E_{QS}}(\lambda;\Omega_{QS}) = \frac{1}{|\lambda||\boldsymbol{y}^*\boldsymbol{x}|}\Bigg\{ & |\boldsymbol{y}^*|Q_d|\boldsymbol{x}| + |\boldsymbol{y}^*|\,|Q_p|\,|C_L\boldsymbol{x}| + |\boldsymbol{y}^*C_L|\,|Q_q|\,|\boldsymbol{x}| \\ & + |\boldsymbol{y}^*|\,|Q_g|\,|C_U\boldsymbol{x}| + |\boldsymbol{y}^*C_U|\,|Q_h|\,|\boldsymbol{x}| \\ & + \sum_{i=2}^{n-1}\left|\boldsymbol{y}^*\begin{bmatrix} 0 & 0 \\ C(i+1:n,1:i-1) & 0\end{bmatrix}\boldsymbol{x}\right|\left|\frac{e_{a_i}}{a_i}\right| \\ & + \sum_{j=2}^{n-1}\left|\boldsymbol{y}^*\begin{bmatrix} 0 & C(1:j-1,j+1:n) \\ 0 & 0\end{bmatrix}\boldsymbol{x}\right|\left|\frac{e_{b_i}}{b_i}\right|\Bigg\},\end{aligned}$$

*where*

$$\begin{aligned}\Omega_{QS} &= (\{p_i\}_{i=2}^n, \{a_i\}_{i=2}^{n-1}, \{q_i\}_{i=1}^{n-1}, \{d_i\}_{i=1}^n, \{g_i\}_{i=1}^{n-1}, \{b_i\}_{i=2}^{n-1}, \{h_i\}_{i=2}^n), \\ E_{QS} &= (\{e_{p_i}\}_{i=2}^n, \{e_{a_i}\}_{i=2}^{n-1}, \{e_{q_i}\}_{i=1}^{n-1}, \{e_{d_i}\}_{i=1}^n, \{e_{g_i}\}_{i=1}^{n-1}, \{e_{b_i}\}_{i=2}^{n-1}, \{e_{h_i}\}_{i=2}^n), \\ Q_d &= \operatorname{diag}(e_{d_1}, e_{d_2}, \ldots, e_{d_n}),\ Q_p = \operatorname{diag}\left(1, \frac{e_{p_2}}{p_2}, \ldots, \frac{e_{p_n}}{p_n}\right), \\ Q_q &= \operatorname{diag}\left(\frac{e_{q_1}}{q_1}, \ldots, \frac{e_{q_{n-1}}}{q_{n-1}}, 1\right),\ Q_g = \operatorname{diag}\left(\frac{e_{g_1}}{g_1}, \ldots, \frac{e_{g_{n-1}}}{g_{n-1}}, 1\right),\end{aligned}$$

$$Q_h = \operatorname{diag}\left(1, \frac{e_{h_2}}{h_2}, \ldots, \frac{e_{h_n}}{h_n}\right),$$

*and each quotient whose denominator is zero must be understood as zero if the numerator is also zero and, otherwise, the zero parameter in the denominator should be formally cancelled out with the same parameter in the corresponding piece of $C$.*

*Proof.* In order to use the expression in Theorem 3.25, we will proceed by parts by calculating the contribution of each subset of parameters as follows.

Derivatives with respect to $\{d_i\}_{i=1}^n$. By using a) in Lemma 4.1 we get:

$$K_d = \sum_{i=1}^{n} \left| \boldsymbol{y}^* \frac{\partial C}{\partial d_i} \boldsymbol{x} \right| e_{d_i} = \sum_{i=1}^{n} \left| \boldsymbol{y}^* \boldsymbol{e}_i \boldsymbol{e}_i^T \boldsymbol{x} \right| e_{d_i} = \sum_{i=1}^{n} |\overline{y}_i| \, |x_i| \, e_{d_i} = |\boldsymbol{y}^*| Q_d |\boldsymbol{x}|.$$

Derivatives with respect to $\{p_i\}_{i=2}^n$. From b) in Lemma 4.1 we have:

$$\begin{aligned} K_p &= \sum_{i=2}^{n} \left| \boldsymbol{y}^* \frac{\partial C}{\partial p_i} \boldsymbol{x} \right| e_{p_i} = \sum_{i=2}^{n} \left| \boldsymbol{y}^* p_i \frac{\partial C}{\partial p_i} \boldsymbol{x} \right| \left| \frac{e_{p_i}}{p_i} \right| = \sum_{i=2}^{n} |\boldsymbol{y}^* \boldsymbol{e}_i C_L(i,:) \boldsymbol{x}| \left| \frac{e_{p_i}}{p_i} \right| \\ &= \sum_{i=2}^{n} |\overline{y}_i C_L(i,:) \boldsymbol{x}| \left| \frac{e_{p_i}}{p_i} \right| = \sum_{i=2}^{n} |\overline{y}_i| \left| \frac{e_{p_i}}{p_i} \right| |C_L(i,:) \boldsymbol{x}| = |\boldsymbol{y}^*| \, |Q_p| \, |C_L \boldsymbol{x}| \, . \end{aligned}$$

Derivatives with respect to $\{a_i\}_{i=2}^{n-1}$. From c) in Lemma 4.1 we have:

$$K_a = \sum_{i=2}^{n-1} \left| \boldsymbol{y}^* \frac{\partial C}{\partial a_i} \boldsymbol{x} \right| e_{a_i} = \sum_{i=2}^{n-1} \left| \boldsymbol{y}^* \begin{bmatrix} 0 & 0 \\ C(i+1:n, 1:i-1) & 0 \end{bmatrix} \boldsymbol{x} \right| \left| \frac{e_{a_i}}{a_i} \right|.$$

Derivatives with respect to $\{q_j\}_{j=1}^{n-1}$. From d) in Lemma 4.1 we have:

$$\begin{aligned} K_q &= \sum_{j=1}^{n-1} \left| \boldsymbol{y}^* \frac{\partial C}{\partial q_j} \boldsymbol{x} \right| e_{q_j} = \sum_{j=1}^{n-1} \left| \boldsymbol{y}^* q_j \frac{\partial C}{\partial q_j} \boldsymbol{x} \right| \left| \frac{e_{q_j}}{q_j} \right| = \sum_{j=1}^{n-1} \left| \boldsymbol{y}^* C_L(:,j) \boldsymbol{e}_j^T \boldsymbol{x} \right| \left| \frac{e_{q_j}}{q_j} \right| \\ &= \sum_{j=1}^{n-1} |\boldsymbol{y}^* C_L(:,j)| \, |x_j| \left| \frac{e_{q_j}}{q_j} \right| = |\boldsymbol{y}^* C_L| \, |Q_q| \, |\boldsymbol{x}| \, . \end{aligned}$$

For the parameters $\{g_i\}_{i=1}^{n-1}, \{b_i\}_{i=2}^{n-1}$, and $\{h_i\}_{i=2}^{n}$, which describe the strictly upper triangular part of $C$, we can proceed analogously and find the contribution to $\operatorname{cond}_{E_{QS}}(\lambda; \Omega_{QS})$ of the derivatives of $C$ with respect to those parameters. The results are the following.

Derivatives with respect to $\{g_i\}_{i=1}^{n-1}$. By using e) in Lemma 4.1 we obtain:

$$K_g = \sum_{i=1}^{n-1} \left| \boldsymbol{y}^* \frac{\partial C}{\partial g_i} \boldsymbol{x} \right| e_{g_i} = |\boldsymbol{y}^*| \, |Q_g| \, |C_U \boldsymbol{x}| \, .$$

Derivatives with respect to $\{b_i\}_{i=2}^{n-1}$. By using f) in Lemma 4.1 we obtain:

$$K_b = \sum_{i=2}^{n-1} \left| \boldsymbol{y}^* \frac{\partial C}{\partial b_i} \boldsymbol{x} \right| e_{b_i} = \sum_{i=2}^{n-1} \left| \boldsymbol{y}^* \begin{bmatrix} 0 & C(1:i-1, i+1:n) \\ 0 & 0 \end{bmatrix} \boldsymbol{x} \right| \left| \frac{e_{b_i}}{b_i} \right|.$$

Derivatives with respect to $\{h_j\}_{j=2}^{n}$. By using g) in Lemma 4.1 we obtain:

$$K_h = \sum_{j=2}^{n} \left| \boldsymbol{y}^* \frac{\partial C}{\partial h_j} \boldsymbol{x} \right| e_{h_j} = |\boldsymbol{y}^* C_U| \, |Q_h| \, |\boldsymbol{x}| \, .$$

We complete this proof by observing that

$$\mathrm{cond}(\lambda; \Omega_{QS}) = \frac{1}{|\lambda||\boldsymbol{y}^* \boldsymbol{x}|} \left( K_d + K_p + K_a + K_q + K_g + K_b + K_h \right).$$

□

Again, as it happens for the condition number $\mathrm{cond}_{E_{QS},\boldsymbol{f}}(A(\Omega_{QS}), \boldsymbol{x})$ in Theorem 4.2 for linear systems, it is easy to see from the expression in Theorem 5.1 that when considering relative perturbations with respect to a general nonnegative vector $E_{QS}$, the eigenvalue condition number $\mathrm{cond}_{E_{QS}}(\lambda; \Omega_{QS})$ depends in general on the quasiseparable representation $\Omega_{QS}$ of the matrix $C$. In fact, this condition number also depends on the ratios between the parameters in the representation and the corresponding tolerances in $E_{QS}$. Therefore, we will restrict ourselves to the case $E_{QS} = |\Omega_{QS}|$ in Theorem 5.2, which is again the most natural election for $E_{QS}$. In this situation we adopt, for brevity in the rest of this thesis, the following notation:

$$\mathrm{cond}(\lambda; \Omega_{QS}) \equiv \mathrm{cond}_{|\Omega_{QS}|}(\lambda; \Omega_{QS}).$$

**Theorem 5.2.** *Let $C \in \mathbb{R}^{n \times n}$ be a $\{1,1\}$-quasiseparable matrix and let us express $C$ as $C = C_L + C_D + C_U$, with $C_L$ strictly lower triangular, $C_D$ diagonal, and $C_U$ strictly upper triangular. Suppose $\lambda \neq 0$ is a simple eigenvalue of $C$ with left and right eigenvectors $\boldsymbol{y}$ and $\boldsymbol{x}$, respectively, and denote by $\Omega_{QS}$ a quasiseparable representation of $C$. Then, the componentwise relative condition number* $\mathrm{cond}(\lambda; \Omega_{QS})$ *of $\lambda$ with respect to $\Omega_{QS}$ is given by the following expression:*

$$\begin{aligned} \mathrm{cond}(\lambda; \Omega_{QS}) = \frac{1}{|\lambda||\boldsymbol{y}^* \boldsymbol{x}|} \Bigg\{ & |\boldsymbol{y}^*||C_D||\boldsymbol{x}| + |\boldsymbol{y}^*||C_L \boldsymbol{x}| + |\boldsymbol{y}^* C_L||\boldsymbol{x}| + |\boldsymbol{y}^*||C_U \boldsymbol{x}| \\ & + |\boldsymbol{y}^* C_U||\boldsymbol{x}| + \sum_{i=2}^{n-1} \left| \boldsymbol{y}^* \begin{bmatrix} 0 & 0 \\ C(i+1:n, 1:i-1) & 0 \end{bmatrix} \boldsymbol{x} \right| \\ & + \sum_{j=2}^{n-1} \left| \boldsymbol{y}^* \begin{bmatrix} 0 & C(1:j-1, j+1:n) \\ 0 & 0 \end{bmatrix} \boldsymbol{x} \right| \Bigg\}. \end{aligned}$$

*Proof.* It follows directly from the expression in Theorem 5.1 for $\text{cond}_{|\Omega_{QS}|}(\lambda;\Omega_{QS})$ by observing that, in this case, we are considering $E_{QS} = |\Omega_{QS}|$ and, therefore, using the notation in Theorem 5.1, the following equalities hold: $Q_d = |C_D|$, $|Q_p| = |Q_q| = |Q_g| = |Q_h| = I$, and $|e_{a_i}/a_i| = |e_{b_i}/b_i| = 1$. □

The explicit formula given in Theorem 5.2 for $\text{cond}(\lambda;\Omega_{QS})$ does not depend on the parameters of the chosen quasiseparable representation; it only depends on the matrix entries, the simple eigenvalue $\lambda$, and the left and right eigenvectors. This important property allows us to state the following proposition.

**Proposition 5.3.** *Let $C \in \mathbb{R}^{n\times n}$ be a {1,1}-quasiseparable matrix and $\lambda \neq 0$ be a simple eigenvalue of $C$. Then, for any two vectors $\Omega_{QS}$ and $\Omega'_{QS}$ of quasiseparable parameters of $C$,*

$$\text{cond}(\lambda;\Omega_{QS}) = \text{cond}(\lambda;\Omega'_{QS}).$$

Another important property for this relative componentwise condition number appears from the natural comparison with the unstructured relative entrywise condition number defined in Definition 3.17 and whose explicit expression is given in Theorem 3.18. This comparison is established in the next proposition.

**Proposition 5.4.** *Let $C$ be a {1,1}-quasiseparable matrix and consider a set of quasiseparable parameters $\Omega_{QS}$ of $C$. Let $\lambda \neq 0$ be a simple eigenvalue of $C$. Then, the following relation holds,*

$$\text{cond}(\lambda;\Omega_{QS}) \leq n\,\text{cond}(\lambda;C).$$

*Proof.* From Theorem 5.2, and standard inequalities of absolute values we get:

$$\begin{aligned}
\text{cond}(\lambda;\Omega_{QS}) &\leq \frac{1}{|\lambda||\boldsymbol{y}^*\boldsymbol{x}|}\Bigg\{|\boldsymbol{y}^*||C_D||\boldsymbol{x}| + |\boldsymbol{y}^*||C_L||\boldsymbol{x}| + |\boldsymbol{y}^*||C_L||\boldsymbol{x}| + |\boldsymbol{y}^*||C_U||\boldsymbol{x}| \\
&\qquad + |\boldsymbol{y}^*||C_U||\boldsymbol{x}| + \sum_{i=2}^{n-1}|\boldsymbol{y}^*||C_L||\boldsymbol{x}| + \sum_{j=2}^{n-1}|\boldsymbol{y}^*||C_U||\boldsymbol{x}|\Bigg\} \\
&\leq \frac{1}{|\lambda||\boldsymbol{y}^*\boldsymbol{x}|}\Big\{|\boldsymbol{y}^*||C_D||\boldsymbol{x}| + n|\boldsymbol{y}^*||C_L||\boldsymbol{x}| + n|\boldsymbol{y}^*||C_U||\boldsymbol{x}|\Big\} \\
&\leq n\frac{|\boldsymbol{y}^*||C||\boldsymbol{x}|}{|\lambda||\boldsymbol{y}^*\boldsymbol{x}|} = n\,\text{cond}(\lambda;C).
\end{aligned}$$

□

According to Proposition 5.4, the structured condition number $\text{cond}(\lambda;\Omega_{QS})$ is smaller than the unstructured condition number $\text{cond}(\lambda;C)$, except for a factor of $n$, but it can be potentially much smaller, as we will see in our numerical experiments (see Section 5.6). The factor $n$ comes from the entries $C_{1n}$ and $C_{n1}$. For instance,

$C_{n1} = p_n a_{n-1} \cdots a_2 q_1$, implies that an entrywise relative perturbation of size $\eta$ over the representation $\Omega_{QS}$ will generate a perturbation on the entry $C_{n1}$ of the matrix $C$ involving $n$ factors of the form $(1+\delta)$, where $\delta$ represents the relative perturbations on the parameters.

On the other hand, the exact expression of $\text{cond}(\lambda; \Omega_{QS})$ deduced in Theorem 5.2 is complicated, especially because of the summations appearing in the last two terms. Surprisingly, these two summations can be removed in order to define the effective condition number introduced in Definition 5.5, which can be used to estimate $\text{cond}(\lambda; \Omega_{QS})$ reliably up to a factor $n$. This is proved in Proposition 5.6.

**Definition 5.5.** *Under the same hypotheses of Theorem 5.2, we define the* effective relative condition number $\text{cond}_{\text{eff}}(\lambda; \Omega_{QS})$ *of* $\lambda$ *with respect to the quasiseparable representation* $\Omega_{QS}$ *of* $C$ *as,*

$$\text{cond}_{\text{eff}}(\lambda; \Omega_{QS}) := \frac{1}{|\lambda||\boldsymbol{y}^*\boldsymbol{x}|} \Big\{ |\boldsymbol{y}^*||C_D||\boldsymbol{x}| + |\boldsymbol{y}^*||C_L\boldsymbol{x}| + |\boldsymbol{y}^* C_L||\boldsymbol{x}| + |\boldsymbol{y}^*||C_U\boldsymbol{x}| + |\boldsymbol{y}^* C_U||\boldsymbol{x}| \Big\}.$$

**Proposition 5.6.** *Let* $C \in \mathbb{R}^{n\times n}$ *be a* $\{1,1\}$*-quasiseparable matrix with a simple eigenvalue* $\lambda \neq 0$ *with left and right eigenvectors* $\boldsymbol{y}$ *and* $\boldsymbol{x}$*, respectively. Let* $\Omega_{QS}$ *be a quasiseparable representation of* $C$*. Then*

$$\text{cond}_{\text{eff}}(\lambda; \Omega_{QS}) \leq \text{cond}(\lambda; \Omega_{QS}) \leq (n-1)\text{cond}_{\text{eff}}(\lambda; \Omega_{QS}).$$

*Proof.* The first inequality is trivial from the definitions of $\text{cond}_{\text{eff}}(\lambda; \Omega_{QS})$ and $\text{cond}(\lambda; \Omega_{QS})$, respectively. On the other hand, note that

$$\begin{aligned}
\left|\boldsymbol{y}^* \begin{bmatrix} 0 & 0 \\ C(i+1:n, 1:i-1) & 0 \end{bmatrix} \boldsymbol{x}\right| \leq & \left|\boldsymbol{y}^* \begin{bmatrix} 0 & 0 \\ C_L(i+1:n, 1:i-1) & C_L(i+1:n, i:n) \end{bmatrix} \boldsymbol{x}\right| \\
& + \left|\boldsymbol{y}^* \begin{bmatrix} 0 & 0 \\ 0 & -C_L(i+1:n, i:n) \end{bmatrix} \boldsymbol{x}\right| \\
\leq & |\boldsymbol{y}^*| \left| \begin{bmatrix} 0 \\ C_L(i+1:n, :) \end{bmatrix} \boldsymbol{x}\right| \\
& + \left|\boldsymbol{y}^* \begin{bmatrix} 0 & 0 \\ 0 & C_L(i+1:n, i:n) \end{bmatrix}\right| |\boldsymbol{x}| \\
\leq & |\boldsymbol{y}^*||C_L\boldsymbol{x}| + |\boldsymbol{y}^* C_L||\boldsymbol{x}|,
\end{aligned}$$

from where we obtain:

$$\sum_{i=2}^{n-1} \left|\boldsymbol{y}^* \begin{bmatrix} 0 & 0 \\ C(i+1:n, 1:i-1) & 0 \end{bmatrix} \boldsymbol{x}\right| \leq (n-2)|\boldsymbol{y}^*||C_L\boldsymbol{x}| + (n-2)|\boldsymbol{y}^* C_L||\boldsymbol{x}|. \quad (5.1)$$

In an analogous way, we can prove that

$$\sum_{j=2}^{n-1}\left|\boldsymbol{y}^*\begin{bmatrix}0 & C(1:j-1,j+1:n)\\ 0 & 0\end{bmatrix}\boldsymbol{x}\right| \leq (n-2)|\boldsymbol{y}^*C_U||\boldsymbol{x}|+(n-2)|\boldsymbol{y}^*||C_U\boldsymbol{x}|. \quad (5.2)$$

Finally, from (5.1) and (5.2) it is straightforward that

$$\mathrm{cond}(\lambda;\Omega_{QS}) \leq \frac{n-1}{|\lambda||\boldsymbol{y}^*\boldsymbol{x}|}\left\{|\boldsymbol{y}^*||C_D||\boldsymbol{x}|+|\boldsymbol{y}^*||C_L\boldsymbol{x}|+|\boldsymbol{y}^*C_L||\boldsymbol{x}|+|\boldsymbol{y}^*||C_U\boldsymbol{x}|+|\boldsymbol{y}^*C_U||\boldsymbol{x}|\right\},$$

which completes the proof. □

Another important property of eigenvalue condition numbers that must be studied is their behavior under diagonal similarities since many algorithms for computing eigenvalues of matrices start by *balancing* the matrix, i.e., by performing a diagonal similarity that for each $i$ makes the norm ($\|\cdot\|_1$, $\|\cdot\|_2$, or $\|\cdot\|_\infty$) of the $i$th row equal to the norm ($\|\cdot\|_1$, $\|\cdot\|_2$, or $\|\cdot\|_\infty$) of the *ith* column (see [52], [38, p. 360-361]). For such purpose, Lemmas 5.7 and 5.8 will be useful for proving Theorem 5.9, which will establish the invariance of $\mathrm{cond}(\lambda;\Omega_{QS})$ under diagonal similarities. In the following, we will denote by $K = \mathrm{diag}\,(k_1, k_2, \cdots, k_n)$ any diagonal matrix $K \in \mathbb{R}^{n\times n}$ such that $K_{ii} = k_i$. For any matrix $C$ and any two ordered sets $\mathcal{I}$ and $\mathcal{J}$, we will denote by $C(\mathcal{I},\mathcal{J})$ the submatrix of $C$ consisting of rows and columns with indices in $\mathcal{I}$ and $\mathcal{J}$, respectively.

**Lemma 5.7.** *Let $K = \mathrm{diag}\,(k_1, k_2, \cdots, k_n)$ be an invertible diagonal matrix, and let $A$, $B \in \mathbb{R}^{n\times n}$ be matrices such that $B = KAK^{-1}$. Then, the following assertions hold.*

*(a) For any two ordered subsets $\mathcal{I}_p = \{i_1, i_2, \cdots, i_p\}$ and $\mathcal{J}_q = \{j_1, j_2, \cdots, j_q\}$ of indices such that $1 \leq i_1 < i_2 < \cdots < i_p \leq n$ and $1 \leq j_1 < j_2 < \cdots < j_q \leq n$, we have that*

$$B\,(\mathcal{I}_p,\mathcal{J}_q) = diag\left(k_{i_1}, k_{i_2}, \ldots, k_{i_p}\right)\cdot\ A\,(\mathcal{I}_p,\mathcal{J}_q)\cdot diag\left(1/k_{j_1}, 1/k_{j_2}, \ldots, 1/k_{j_q}\right).$$

*(b) For any matrix $C \in \mathbb{R}^{n\times n}$, let us denote by $\tilde{C}_{i_{1:p},j_{1:q}} \in \mathbb{R}^{n\times n}$ a matrix such that $\tilde{C}_{i_{1:p},j_{1:q}}\,(\mathcal{I}_p,\mathcal{J}_q) = C\,(\mathcal{I}_p,\mathcal{J}_q)$, and $\tilde{C}_{i_{1:p},j_{1:q}}\,(i,j) = 0$, for any other entry. Then*

$$\tilde{B}_{i_{1:p},j_{1:q}} = K\tilde{A}_{i_{1:p},j_{1:q}}K^{-1}.$$

*(c) If we decompose the matrices $A = A_L+A_D+A_U$ and $B = B_L+B_D+B_U$, where $A_L$ and $B_L$ are strictly lower triangular matrices, $A_D$ and $B_D$ are diagonal matrices, and $A_U$ and $B_U$ are strictly upper triangular matrices, then*

$$B_L = KA_LK^{-1},\ B_D = KA_DK^{-1} = A_D,\ \text{and } B_U = KA_UK^{-1}.$$

*(d) A vector $\boldsymbol{x}_A \in \mathbb{R}^{n\times 1}$ is a right eigenvector of $A$ associated to the eigenvalue $\lambda$ if and only if the vector $\boldsymbol{x}_B = K\boldsymbol{x}_A$ is a right eigenvector of $B$ associated to $\lambda$. Similarly, a vector $\boldsymbol{y}_A \in \mathbb{R}^{n\times 1}$ is a left eigenvector of $A$ associated to $\lambda$ if and only if the vector $\boldsymbol{y}_B = (\boldsymbol{y}_A^* K^{-1})^*$ is a left eigenvector of $B$ associated to $\lambda$.*

*Proof.* The proofs of (a), (b), and (c) are straightforward. So, we only prove part (d). The result follows from the equivalences:

$$\begin{aligned} A\boldsymbol{x}_A = \lambda\boldsymbol{x}_A \Leftrightarrow \;& K^{-1}BK\boldsymbol{x}_A = \lambda\boldsymbol{x}_A \Leftrightarrow B\left(K\boldsymbol{x}_A\right) = \lambda\left(K\boldsymbol{x}_A\right), \text{ and} \\ \boldsymbol{y}_A^* A = \lambda\boldsymbol{y}_A^* \Leftrightarrow \;& \boldsymbol{y}_A^* K^{-1}BK = \lambda\boldsymbol{y}_A^* \Leftrightarrow \left(\boldsymbol{y}_A^* K^{-1}\right) B = \lambda\left(\boldsymbol{y}_A^* K^{-1}\right). \end{aligned}$$

□

**Lemma 5.8.** *Let $K = \operatorname{diag}(k_1, k_2, \cdots, k_n)$ be an invertible diagonal matrix and $C \in \mathbb{R}^{n\times n}$ be a $\{1,1\}$-quasiseparable matrix. Then, the following two assertions hold.*

*(a) The matrix $KCK^{-1}$ is also a $\{1,1\}$-quasiseparable matrix.*

*(b) If the vector of parameters*

$$\Omega_{QS} = (\{p_i\}_{i=2}^n, \{a_i\}_{i=2}^{n-1}, \{q_i\}_{i=1}^{n-1}, \{d_i\}_{i=1}^n, \{g_i\}_{i=1}^{n-1}, \{b_i\}_{i=2}^{n-1}, \{h_i\}_{i=2}^n)$$

*is a quasiseparable representation of $C$, then the vector of parameters*

$$\Omega'_{QS} = \left(\{k_i p_i\}_{i=2}^n, \{a_i\}_{i=2}^{n-1}, \{q_i/k_i\}_{i=1}^{n-1}, \; \{d_i\}_{i=1}^n, \{k_i g_i\}_{i=1}^{n-1}, \{b_i\}_{i=2}^{n-1}, \{h_i/k_i\}_{i=2}^n\right),$$

*is a quasiseparable representation of $KCK^{-1}$.*

*Proof.* (a) It follows from $(a)$ in Lemma 5.7 that for any two subsets $\{i_1, i_2, \cdots, i_p\}$ and $\{j_1, j_2, \cdots, j_q\}$ of indices such that $1 \le i_1 < i_2 < \cdots < i_p \le n$ and $1 \le j_1 < j_2 < \cdots < j_q \le n$, we have that

$$\begin{aligned} &\operatorname{rank} KCK^{-1}\left(\{i_1, i_2, \cdots, i_p\}, \{j_1, j_2, \cdots, j_q\}\right) \\ &= \operatorname{rank} C\left(\{i_1, i_2, \cdots, i_p\}, \{j_1, j_2, \cdots, j_q\}\right), \end{aligned}$$

and the result follows from Definition 2.9.

(b) It follows from Theorem 2.18 and from $KCK^{-1}(i,j) = k_i C(i,j)\frac{1}{k_j}$.

□

**Theorem 5.9.** *Let $C \in \mathbb{R}^{n\times n}$ be $\{1,1\}$-quasiseparable, $\lambda \neq 0$ be a simple eigenvalue of $C$, $K \in \mathbb{R}^{n\times n}$ be diagonal and nonsingular, $\Omega_{QS}$ be any vector of quasiseparable parameters of $C$, and $\Omega'_{QS}$ be any vector of quasiseparable parameters of $KCK^{-1}$. Then*

$$\operatorname{cond}(\lambda, C; \Omega_{QS}) = \operatorname{cond}(\lambda, KCK^{-1}; \Omega'_{QS}).$$

*Proof.* Note first that for any two matrices $A \in \mathbb{R}^{n\times n}$ and $B \in \mathbb{R}^{n\times n}$ we have that $|AK|\,|B| = |A|\,|K|\,|B| = |A|\,|KB|$, since $K$ is diagonal. Let us consider $B = KCK^{-1}$ and let us analyze the expression given in Theorem 5.2 for $\mathrm{cond}\,(\lambda, C; \Omega_{QS})$, term by term. Using Lemma 5.7 we see that:

1) $|\boldsymbol{y}_C^* \boldsymbol{x}_C| = |\boldsymbol{y}_C^* K^{-1} K \boldsymbol{x}_C| = |\boldsymbol{y}_B^* \boldsymbol{x}_B|\,,$

2) $|\boldsymbol{y}_C^*|\,|C_D|\,|\boldsymbol{x}_C| = |\boldsymbol{y}_B^* K|\,|C_D|\,|K^{-1}\boldsymbol{x}_B| = |\boldsymbol{y}_B^*|\,|KC_DK^{-1}|\,|\boldsymbol{x}_B| = |\boldsymbol{y}_B^*|\,|B_D|\,|\boldsymbol{x}_B|\,,$

3) $|\boldsymbol{y}_C^*|\,|C_L\boldsymbol{x}_C| = |\boldsymbol{y}_B^* K|\,|C_L K^{-1}\boldsymbol{x}_B| = |\boldsymbol{y}_B^*|\,|KC_L K^{-1}\boldsymbol{x}_B| = |\boldsymbol{y}_B^*|\,|B_L\boldsymbol{x}_B|\,,$

4) $|\boldsymbol{y}_C^* C_L|\,|\boldsymbol{x}_C| = |\boldsymbol{y}_B^* KC_L|\,|K^{-1}\boldsymbol{x}_B| = |\boldsymbol{y}_B^* KC_L K^{-1}|\,|\boldsymbol{x}_B| = |\boldsymbol{y}_B^* B_L|\,|\boldsymbol{x}_B|\,,$

5) $|\boldsymbol{y}_C^*|\,|C_U\boldsymbol{x}_C| = |\boldsymbol{y}_B^* K|\,|C_U K^{-1}\boldsymbol{x}_B| = |\boldsymbol{y}_B^*|\,|KC_U K^{-1}\boldsymbol{x}_B| = |\boldsymbol{y}_B^*|\,|B_U\boldsymbol{x}_B|\,,$

6) $|\boldsymbol{y}_C^* C_U|\,|\boldsymbol{x}_C| = |\boldsymbol{y}_B^* KC_U|\,|K^{-1}\boldsymbol{x}_B| = |\boldsymbol{y}_B^* KC_U K^{-1}|\,|\boldsymbol{x}_B| = |\boldsymbol{y}_B^* B_U|\,|\boldsymbol{x}_B|\,,$

7) $\left|\boldsymbol{y}_C^* \begin{bmatrix} 0 & 0 \\ C(i+1:n, 1:i-1) & 0 \end{bmatrix} \boldsymbol{x}_C\right| = \left|\boldsymbol{y}_B^* \begin{bmatrix} 0 & 0 \\ B(i+1:n, 1:i-1) & 0 \end{bmatrix} \boldsymbol{x}_B\right|,$

8) $\left|\boldsymbol{y}_C^* \begin{bmatrix} 0 & C(1:j-1, j+1:n) \\ 0 & 0 \end{bmatrix} \boldsymbol{x}_C\right| = \left|\boldsymbol{y}_B^* \begin{bmatrix} 0 & C(1:j-1, j+1:n) \\ 0 & 0 \end{bmatrix} \boldsymbol{x}_B\right|.$

The result follows from Theorem 5.2, and from items 1) through 8) above. □

## 5.2 Fast computation of the eigenvalue condition number in the quasiseparable representation

The main contribution of this section is that, via Proposition 5.10, we will provide an algorithm for computing the eigenvalue condition number $\mathrm{cond}\,(\lambda; \Omega_{QS})$ for any simple eigenvalue of any $\{1,1\}$-quasiseparable matrix $C$ of size $n \times n$ in $\mathcal{O}(n)$ operations. Taking into account that fast algorithms for computing all the eigenvalues of a quasiseparable matrix cost $\mathcal{O}(n^2)$ flops [62], our result allows us to compute the condition numbers of all the eigenvalues of a quasiseparable matrix also in $\mathcal{O}(n^2)$ flops.

We remark that the difficulty of computing $\mathrm{cond}\,(\lambda; \Omega_{QS})$ fast comes mainly from the terms:

$$\sum_{i=2}^{n-1} \left|\boldsymbol{y}^* \begin{bmatrix} 0 & 0 \\ C(i+1:n, 1:i-1) & 0 \end{bmatrix} \boldsymbol{x}\right| \text{ and } \sum_{j=2}^{n-1} \left|\boldsymbol{y}^* \begin{bmatrix} 0 & C(1:j-1, j+1:n) \\ 0 & 0 \end{bmatrix} \boldsymbol{x}\right|,$$

that appear in the formula given in Theorem 5.2. Note that the computation of these terms may be avoided, as a consequence of Proposition 5.6, if we estimate $\mathrm{cond}\,(\lambda; \Omega_{QS})$ via $\mathrm{cond}_{\mathrm{eff}}\,(\lambda; \Omega_{QS})$.

**Proposition 5.10.** *Let $C \in \mathbb{R}^{n\times n}$ be a $\{1,1\}$-quasiseparable matrix with a simple eigenvalue $\lambda \neq 0$ with left eigenvector $\boldsymbol{y}$ and right eigenvector $\boldsymbol{x}$, and assume that $\lambda, \boldsymbol{x}, \boldsymbol{y}$, and a quasiseparable representation $\Omega_{QS}$ of $C$ are all known. Then,* $\operatorname{cond}(\lambda; \Omega_{QS})$ *can be computed in* $42n - 66$ *flops.*

*Proof.* This proof consists of giving an algorithm for calculating $\operatorname{cond}(\lambda; \Omega_{QS})$. We will count the number of flops needed for calculating the expression of $\operatorname{cond}(\lambda; \Omega_{QS})$, term by term as follows.

(a) The factor $|\lambda|\,|\boldsymbol{y}^*\boldsymbol{x}| = \left|\lambda \sum_{i=1}^{n} \overline{y}_i x_i\right|$ requires $2n$ flops.

(b) Since every product $\overline{y}_i x_i$ has already been calculated in (a), the term

$$|\boldsymbol{y}^*|\,|C_D|\,|\boldsymbol{x}| = \sum_{i=1}^{n} |d_i|\,|\overline{y}_i x_i|,$$

can be calculated in $2n - 1$ flops.

(c) For the term $|\boldsymbol{y}^*|\,|C_L\boldsymbol{x}|$, we will calculate the products $C_L\boldsymbol{x}$ and $C_L^{(-2)}\boldsymbol{x}$ simultaneously, where $C_L^{(-2)}$ denotes the matrix that is obtained from $C_L$ by setting to zero the entries $(C_L)_{i+1,i}$ for $i = 1, 2, \ldots, n-1$, and that will be used later for calculating the term in (g). For simplicity, let us denote the vectors $\mathrm{wlx} = C_L^{(-2)}\boldsymbol{x}$ and $\mathrm{zlx} = C_L\boldsymbol{x}$. Notice that $\mathrm{wlx}_1 = \mathrm{wlx}_2 = 0$ and $\mathrm{zlx}_1 = 0$. The fast method for calculating $C_L\boldsymbol{x}$ is given by the following algorithm.

**Routine 1** (Computes $\mathrm{zlx} = C_L\boldsymbol{x}$ and $\mathrm{wlx} = C_L^{(-2)}\boldsymbol{x}$ taking as inputs the quasiseparable parameters $\{p_i\}_{i=2}^{n}, \{a_i\}_{i=2}^{n-1}, \{q_i\}_{i=1}^{n-1}$, and the entries $\{x_i\}_{i=1}^{n}$ of $\boldsymbol{x}$.)

$\mathrm{zlx}_1 = \mathrm{wlx}_1 = \mathrm{wlx}_2 = 0$
$\mathrm{tlx1} = q_1 \cdot x_1$
$\mathrm{zlx}_2 = p_2 \cdot \mathrm{tlx1}$

`for` $i = 3 : n$

  $\mathrm{tlx2} = a_{i-1} \cdot \mathrm{tlx1}$
  $\mathrm{wlx}_i = p_i \cdot \mathrm{tlx2}$
  $\mathrm{tlx1} = \mathrm{tlx2} + q_{i-1} \cdot x_{i-1}$
  $\mathrm{zlx}_i = p_i \cdot \mathrm{tlx1}$

`endfor`

The fact that Routine 1 indeed computes $\mathrm{zlx} = C_L\boldsymbol{x}$ and $\mathrm{wlx} = C_L^{(-2)}\boldsymbol{x}$ can be proved easily by induction. Observe that Routine 1 uses the temporary variables tlx1 and tlx2, in addition to the entries of the vectors zlx and wlx.

We warn the reader that these variables will be used also in Algorithm 1 for computing $\mathrm{cond}(\lambda;\Omega_{QS})$, which is developed and presented in Section 5.2.1, and are described in the table right before Algorithm 1. From Routine 1, we see that the cost of calculating $C_L\boldsymbol{x}$ and $C_L^{(-2)}\boldsymbol{x}$ is $5n-8$ flops, and taking into account that $(C_L\boldsymbol{x})_1 = 0$ it is straightforward that the cost of calculating $|\boldsymbol{y}^*|\,|C_L\boldsymbol{x}|$ and $|C_L^{(-2)}\boldsymbol{x}|$ simultaneously is $7n-11$ flops.

(d) For the term $|\boldsymbol{y}^*C_L|\,|\boldsymbol{x}|$, we can use a similar procedure to that in Routine 1 to compute $\mathrm{zly} = \boldsymbol{y}^*C_L$ and $\mathrm{wly} = \boldsymbol{y}^*C_L^{(-2)}$ by starting with the last components of these two row vectors. As before, this can be done at the cost of $5n-8$ flops, and then, the cost of calculating $|\boldsymbol{y}^*C_L|\,|\boldsymbol{x}|$ and $\boldsymbol{y}^*C_L^{(-2)}$ is $7n-11$ flops.

(e) For the term $|\boldsymbol{y}^*|\,|C_U\boldsymbol{x}|$, we can also calculate $\mathrm{zux} = C_U\boldsymbol{x}$ and $\mathrm{wux} = C_U^{(+2)}\boldsymbol{x}$ (where $C_U^{(+2)}$ denotes the matrix that is obtained from $C_U$ by setting to zero the entries $(C_U)_{i,i+1}$) by using an analogous process to that in Routine 1. Note that calculating $C_U\boldsymbol{x}$ is similar to computing $\boldsymbol{y}^*C_L$ since $(\boldsymbol{y}^*C_L)^* = C_L^T\boldsymbol{y}$. Therefore, the cost of calculating $|\boldsymbol{y}^*|\,|C_U\boldsymbol{x}|$ and $C_U^{(+2)}\boldsymbol{x}$ is of $7n-11$ flops.

(f) The term $|\boldsymbol{y}^*C_U|\,|\boldsymbol{x}|$, can also be computed simultaneously with $\mathrm{wuy} = \boldsymbol{y}^*C_U^{(+2)}$ at a cost of $7n-11$ flops, since the computation of $\mathrm{zuy} = \boldsymbol{y}^*C_U$ is similar to the calculation of $C_L\boldsymbol{x}$, since $(\boldsymbol{y}^*C_U)^* = C_U^T\boldsymbol{y}$.

(g) Denote $\alpha_i = \boldsymbol{y}^*\begin{bmatrix} 0 & 0 \\ C(i+1:n,1:i-1) & 0 \end{bmatrix}\boldsymbol{x}$, and note that

$$\begin{aligned}\alpha_i =& \boldsymbol{y}^*\begin{bmatrix} 0 & 0 \\ C(i+2:n,1:i) & 0 \end{bmatrix}\boldsymbol{x} + \boldsymbol{y}^*\begin{bmatrix} 0 & 0 \\ C(i+1,1:i-1) & 0 \\ 0 & 0 \end{bmatrix}\boldsymbol{x} \\ &- \boldsymbol{y}^*\begin{bmatrix} 0 & 0 & 0 \\ 0 & C(i+2:n,i) & 0 \end{bmatrix}\boldsymbol{x}.\end{aligned}$$

We have obtained the following recursive equation,

$$\alpha_i = \alpha_{i+1} + \overline{y}_{i+1}\left(C_L^{(-2)}\boldsymbol{x}\right)_{i+1} - \left(\boldsymbol{y}^*C_L^{(-2)}\right)_i x_i. \tag{5.3}$$

Recall that $C_L^{(-2)}\boldsymbol{x}$ and $\boldsymbol{y}^*C_L^{(-2)}$ were already calculated in (c) and (d), respectively. Then, the recurrence above is completed with the following fact:

$$\alpha_{n-1} = \boldsymbol{y}^*\begin{bmatrix} 0 & 0 \\ C(n,1:n-2) & 0 \end{bmatrix}\boldsymbol{x} = \overline{y}_n\left(C_L^{(-2)}\boldsymbol{x}\right)_n.$$

Therefore, we can calculate the set $\{\alpha_{n-1},\alpha_{n-2},\cdots,\alpha_2\}$ in $4n-11$ flops and the cost of calculating $\sum_{i=2}^{n-1}|\alpha_i|$, is $5n-14$ flops.

(h) Analogously to the term above, the last sum in the formula for cond $(\lambda; \Omega_{QS})$ can be computed in $5n - 14$ flops via a similar recurrence relation for

$$\beta_j = \boldsymbol{y}^* \begin{bmatrix} 0 & C(1:j-1, j+1:n) \\ 0 & 0 \end{bmatrix} \boldsymbol{x},$$

as we see below

$$\beta_j = \beta_{j-1} + \overline{y}_{j-1} \left( C_U^{(+2)} \boldsymbol{x} \right)_{j-1} - \left( \boldsymbol{y}^* C_U^{(+2)} \right)_j x_j, \quad \text{and}$$
$$\beta_2 = \overline{y}_1 \left( C_U^{(+2)} \boldsymbol{x} \right)_1 .$$

Finally, by summing all the costs in $(a)$-$(h)$, and considering the expression for cond $(\lambda; \Omega_{QS})$, we conclude that the cost of computing it is $42n - 66$ flops. □

### 5.2.1 Pseudocode for computing cond $(\lambda; \Omega_{QS})$ fast

Based on the proof of Proposition 5.10 one can construct an algorithm for computing cond $(\lambda; \Omega_{QS})$ fast. In this section we present, in Algorithm 1, the pseudocode for implementing such computations. If the reader is not interested in technical details, this section may be omitted in a first reading. In the pseudocode we present, we will use the notation of the proof of Proposition 5.10. We will also use the functions 'zeros' (zeros$(m, n)$ returns an $m \times n$ matrix with zero entries), 'sum' (sum$(\boldsymbol{z})$ returns the sum of the entries of the input vector $\boldsymbol{z}$) and 'conj' (conj$(\boldsymbol{z})$ returns an array of the same size of $\boldsymbol{z}$ such that its entries are the conjugates of the respective entries of $\boldsymbol{z}$) from Matlab. We denote by $\boldsymbol{d}$ the column vector of size $n$, such that $\boldsymbol{d}_i = d_i$, where $d_i$ is the $i$-th diagonal entry of $C$. In the following table, we briefly describe the different variables that appear in Algorithm 1.

| **Variable** | **Description** |
|---|---|
| zlx | zlx $= C_L \boldsymbol{x}$ |
| wlx | wlx $= C_L^{(-2)} \boldsymbol{x}$ |
| zly | zly $= \boldsymbol{y}^* C_L$ |
| wly | wly $= \boldsymbol{y}^* C_L^{(-2)}$ |
| zux | zux $= C_U \boldsymbol{x}$ |
| wux | wux $= C_U^{(+2)} \boldsymbol{x}$ |
| zuy | zuy $= \boldsymbol{y}^* C_U$ |
| wuy | wuy $= \boldsymbol{y}^* C_U^{(+2)}$ |
| tlx1 | temporary variable introduced for the fast computation of zlx $= C_L \boldsymbol{x}$ |
| tly1 | temporary variable introduced for the fast computation of zly $= \boldsymbol{y}^* C_L$ |
| tux1 | temporary variable introduced for the fast computation of zux $= C_U \boldsymbol{x}$ |
| tuy1 | temporary variable introduced for the fast computation of zuy $= \boldsymbol{y}^* C_U$ |
| tlx2 | temporary variable introduced for the fast computation of wlx $= C_L^{(-2)} \boldsymbol{x}$ |
| tly2 | temporary variable introduced for the fast computation of wly $= \boldsymbol{y}^* C_L^{(-2)}$ |
| tux2 | temporary variable introduced for the fast computation of wux $= C_U^{(+2)} \boldsymbol{x}$ |
| tuy2 | temporary variable introduced for the fast computation of wuy $= \boldsymbol{y}^* C_U^{(+2)}$ |
| $\alpha$ | vector such that $\alpha(i) = \alpha_i$ with $\alpha_i$ as in (g) in the proof of Proposition 5.10 |
| $\beta$ | vector such that $\beta(i) = \beta_i$ with $\beta_i$ as in (h) in the proof of Proposition 5.10 |

**Algorithm 1** Fast computation of the eigenvalue condition number $\mathrm{cond}\,(\lambda;\Omega_{QS})$

**Input:** quasiseparable parameters $\{p_i\}_{i=2}^{n}, \{a_i\}_{i=2}^{n-1}, \{q_i\}_{i=1}^{n-1}, \{d_i\}_{i=1}^{n}, \{g_i\}_{i=1}^{n-1}$, $\{b_i\}_{i=2}^{n-1}, \{h_i\}_{i=2}^{n}$, the eigenvalue $\lambda$ of $C$, the respective left and right eigenvectors $\boldsymbol{y}$ and $\boldsymbol{x}$.

**Set** zlx= zeros $(n,1)$, wlx= zeros $(n,1)$, zly= zeros $(1,n)$, wly= zeros $(1,n)$, zux = zeros $(n,1)$, wux = zeros $(n,1)$, zuy = zeros $(1,n)$, wuy = zeros $(1,n)$;
$\mathrm{tlx1} = q_1 \cdot x_1$, $\mathrm{zlx}_2 = p_2 \cdot \mathrm{tlx1}$, $\mathrm{tly1} = \overline{y}_n \cdot p_n$, $\mathrm{zly}_{n-1} = q_{n-1} \cdot \mathrm{tly1}$,
$\mathrm{tux1} = x_n \cdot h_n$, $\mathrm{zux}_{n-1} = g_{n-1} \cdot \mathrm{tux1}$, $\mathrm{tuy1} = g_1 \cdot \overline{y}_1$, $\mathrm{zuy}_2 = h_2 \cdot \mathrm{tuy1}$.
**for** i=3 **to** n **do**
  $\mathrm{tlx2} = a_{i-1} \cdot \mathrm{tlx1}$, $\mathrm{wlx}_i = p_i \cdot \mathrm{tlx2}$, $\mathrm{tlx1} = \mathrm{tlx2} + q_{i-1} \cdot x_{i-1}$, $\mathrm{zlx}_i = p_i \cdot \mathrm{tlx1}$;
  $\mathrm{tly2} = a_{n-i+1} \cdot \mathrm{tly1}$, $\mathrm{wly}_{n-i+1} = q_{n-i+1} \cdot \mathrm{tly2}$, $\mathrm{tly1} = \mathrm{tly2} + p_{n-i+2} \cdot \overline{y}_{n-i+2}$, $\mathrm{zly}_{n-i+1} = q_{n-i+1} \cdot \mathrm{tly1}$;
  $\mathrm{tux2} = b_{n-i+1} \cdot \mathrm{tux1}$, $\mathrm{wux}_{n-i+1} = g_{n-i+1} \cdot \mathrm{tux2}$, $\mathrm{tux1} = \mathrm{tux2} + h_{n-i+2} \cdot x_{n-i+2}$, $\mathrm{zux}_{n-i+1} = g_{n-i+1} \cdot \mathrm{tux1}$;
  $\mathrm{tuy2} = b_{i-1} \cdot \mathrm{tuy1}$, $\mathrm{wuy}_i = h_i \cdot \mathrm{tuy2}$, $\mathrm{tuy1} = \mathrm{tuy2} + g_{i-1} \cdot \overline{y}_{i-1}$, $\mathrm{zuy}_i = h_i \cdot \mathrm{tuy1}$.
**end for**
**Set** $\alpha = \mathrm{zeros}\,(1,n)$, $\beta = \mathrm{zeros}\,(1,n)$, $\alpha_{n-1} = \overline{y}_n \cdot \mathrm{wlx}_n$, $\beta_2 = \overline{y}_1 \cdot \mathrm{wux}_1$.
**for** i=3 **to** n-1 **do**
  $\alpha_{n-i+1} = \alpha_{n-i+2} + \overline{y}_{n-i+2} \cdot \mathrm{wlx}_{n-i+2} - \mathrm{wly}_{n-i+1} \cdot x_{n-i+1}$;
  $\beta_i = \beta_{i-1} + \overline{y}_{i-1} \cdot \mathrm{wux}_{i-1} - \mathrm{wuy}_i \cdot x_i$;
**end for**
**Set** $\mathrm{yx} = \mathrm{conj}(\boldsymbol{y}) \,.*\, \boldsymbol{x}$;
$\mathrm{cond}\,(\lambda;\Omega_{QS}) = \big( |\boldsymbol{d}'| \cdot |\mathrm{yx}| + |\boldsymbol{y}'| \cdot |\mathrm{zlx}| + |\mathrm{zly}| \cdot |\boldsymbol{x}| + |\boldsymbol{y}'| \cdot |\mathrm{zux}| + |\mathrm{zuy}| \cdot |\boldsymbol{x}| + \mathrm{sum}\,(|\alpha|) + \mathrm{sum}(|\beta|)\big) / (|\lambda| \cdot |\mathrm{sum}(\mathrm{yx})|)$.

**Output:** $\mathrm{cond}\,(\lambda;\Omega_{QS})$.

We remark that the temporary variables tlx1, tlx2, tly1, tly2, tux1, tux2, tuy1, and tuy2 have been introduced in order to save operations.

## 5.3 Eigenvalue condition numbers for $\{1,1\}$-quasiseparable matrices in the Givens-vector representation via tangents

This section has, partially, a similar structure to that in Section 5.1. The next theorem is the main result of this section and it presents an explicit expression for the componentwise eigenvalue condition number in the Givens-vector representation via tangents introduced in Definition 4.8. The proof of Theorem 5.11 follows from the key Theorem 3.25.

**Theorem 5.11.** *Let $C \in \mathbb{R}^{n\times n}$ be a $\{1,1\}$-quasiseparable matrix, let $\lambda \neq 0$ be a simple eigenvalue of $C$ with right eigenvector $\boldsymbol{x}$ and left eigenvector $\boldsymbol{y}$, and let*

*$C = C_L + C_D + C_U$, with $C_L$ strictly lower triangular, $C_D$ diagonal, and $C_U$ strictly upper triangular. Denote by $\Omega_{GV}$ the tangent-Givens-vector representation of $C$ and let $0 \leq E_{GV} \in \mathbb{R}^{5n-6}$. Then*

$$\begin{aligned}\mathrm{cond}_{E_{GV}}(\lambda;\Omega_{GV}) = \frac{1}{|\lambda||\boldsymbol{y}^*\boldsymbol{x}|}\Bigg\{ & |\boldsymbol{y}^*|Q_d|\boldsymbol{x}| + |\boldsymbol{y}^*C_L|\,|Q_v|\,|\boldsymbol{x}| + |\boldsymbol{y}^*|\,|Q_e|\,|C_U\boldsymbol{x}| \\ & + \sum_{i=2}^{n-1}\left|\boldsymbol{y}^*\begin{bmatrix} 0 & 0 \\ -s_i^2\,C(i,1:i-1) & 0 \\ c_i^2\,C(i+1:n,1:i-1) & 0 \end{bmatrix}\boldsymbol{x}\right|\left|\frac{e_{l_i}}{l_i}\right| \\ & + \sum_{j=2}^{n-1}\left|\boldsymbol{y}^*\begin{bmatrix} 0 & -t_j^2C(1:j-1,j) & r_j^2C(1:j-1,j+1:n) \\ 0 & 0 & 0 \end{bmatrix}\boldsymbol{x}\right|\left|\frac{e_{u_j}}{u_j}\right|\Bigg\}.\end{aligned}$$

*where*

$$\begin{aligned}\Omega_{GV} &= \left(\{l_i\}_{i=2}^{n-1}, \{v_i\}_{i=1}^{n-1}, \{d_i\}_{i=1}^{n}, \{e_i\}_{i=1}^{n-1}, \{u_i\}_{i=2}^{n-1}\right), \text{ as in Definition 4.8,} \\ E_{GV} &= \left(\{e_{l_i}\}_{i=2}^{n-1}, \{e_{v_i}\}_{i=1}^{n-1}, \{e_{d_i}\}_{i=1}^{n}, \{e_{e_i}\}_{i=1}^{n-1}, \{e_{u_i}\}_{i=2}^{n-1}\right), \\ Q_d &= \mathrm{diag}\,(e_{d_1},\ldots,e_{d_n}),\ Q_v = \mathrm{diag}\left(\frac{e_{v_1}}{v_1},\ldots,\frac{e_{v_{n-1}}}{v_{n-1}},1\right), \\ Q_e &= \mathrm{diag}\left(\frac{e_{e_1}}{e_1},\ldots,\frac{e_{e_{n-1}}}{e_{n-1}},1\right),\end{aligned}$$

*and each quotient whose denominator is zero must be understood as zero if the numerator is also zero and, otherwise, the zero parameter in the denominator should be formally cancelled out with the same parameter in the corresponding piece of $A$.*

*Proof.* As in the proof of Theorem 5.1, we will proceed term by term, using the formula (3.18) in Theorem 3.25.
Derivatives with respect to the parameters $\{l_i\}_{i=2}^{n-1}$: Recall first that, from a) in Lemma 4.9, we have

$$l_i\frac{\partial C}{\partial l_i} = \begin{bmatrix} 0 & 0 \\ -s_i^2\,C(i,1:i-1) & 0 \\ c_i^2\,C(i+1:n,1:i-1) & 0 \end{bmatrix},\ i = 2:n-1.$$

Derivatives with respect to the parameters $\{v_j\}_{j=1}^{n-1}$: Note that

$$v_j\frac{\partial C}{\partial v_j} = \begin{bmatrix} 0 & \cdots & 0 & 0 & 0 & \cdots & 0 \\ 0 & \cdots & 0 & C(j+1:n,j) & 0 & \cdots & 0 \end{bmatrix} = C_L(:,j)\boldsymbol{e}_j^T.$$

Derivatives with respect to the parameters $\{d_i\}_{i=1}^{n}$:

$$\frac{\partial C}{\partial d_i} = \boldsymbol{e}_i\boldsymbol{e}_i^T.$$

Derivatives with respect to the parameters $\{e_i\}_{i=1}^{n-1}$:

$$e_i \frac{\partial C}{\partial e_i} = \boldsymbol{e}_i C_U(i,:).$$

Derivatives with respect to the parameters $\{u_j\}_{j=2}^{n-1}$: Recall that, from b) in Lemma 4.9, we have

$$u_j \frac{\partial C}{\partial u_j} = \begin{bmatrix} 0 & -t_j^2 C(1:j-1,j) & r_j^2 C(1:j-1,j+1:n) \\ 0 & 0 & 0 \end{bmatrix}, \ j = 2:n-1.$$

Again, as in the proof of Theorem 5.1, we consider the sums involving the partial derivatives with respect to each subset of parameters:

- $k_l = \displaystyle\sum_{i=2}^{n-1} \left| \boldsymbol{y}^* \frac{\partial C}{\partial l_i} \boldsymbol{x} \right| e_{l_i} = \sum_{i=2}^{n-1} \left| \boldsymbol{y}^* \begin{bmatrix} 0 & 0 \\ -s_i^2\, C(i,1:i-1) & 0 \\ c_i^2\, C(i+1:n,1:i-1) & 0 \end{bmatrix} \boldsymbol{x} \right| \left| \frac{e_{l_i}}{l_i} \right|;$
- $k_v = \displaystyle\sum_{j=1}^{n-1} \left| \boldsymbol{y}^* \frac{\partial C}{\partial v_j} \boldsymbol{x} \right| e_{v_j} = \sum_{j=1}^{n-1} |\boldsymbol{y}^* C_L(:,j) x_j| \left| \frac{e_{v_j}}{v_j} \right| = |\boldsymbol{y}^* C_L| \, |Q_v| \, |\boldsymbol{x}|;$
- $k_d = \displaystyle\sum_{i=1}^{n} \left| \boldsymbol{y}^* \frac{\partial C}{\partial d_i} \boldsymbol{x} \right| e_{d_i} = \sum_{i=1}^{n} |\overline{y}_i x_i| \, e_{d_i} = |\boldsymbol{y}^*| \, Q_d \, |\boldsymbol{x}|;$
- $k_e = \displaystyle\sum_{i=1}^{n-1} \left| \boldsymbol{y}^* \frac{\partial C}{\partial e_i} \boldsymbol{x} \right| e_{e_i} = \sum_{i=1}^{n-1} |\overline{y}_i C_U(i,:) \boldsymbol{x}| \left| \frac{e_{e_i}}{e_i} \right| = |\boldsymbol{y}^*| \, |Q_e| \, |C_U \boldsymbol{x}|;$
- $k_u = \displaystyle\sum_{j=2}^{n-1} \left| \boldsymbol{y}^* \frac{\partial C}{\partial u_j} \boldsymbol{x} \right| e_{u_j} =$
  $\displaystyle\sum_{j=2}^{n-1} \left| \boldsymbol{y}^* \begin{bmatrix} 0 & -t_j^2 C(1:j-1,j) & r_j^2 C(1:j-1,j+1:n) \\ 0 & 0 & 0 \end{bmatrix} \boldsymbol{x} \right| \left| \frac{e_{u_j}}{u_j} \right|.$

The proof is concluded by observing that, from Theorem 3.25, we have

$$\mathrm{cond}_{E_{GV}}(\lambda; \Omega_{GV}) = \frac{1}{|\lambda(\boldsymbol{y}^* \boldsymbol{x})|} \left( k_l + k_v + k_d + k_e + k_u \right).$$

□

Through the rest of this chapter, we will consider the most natural choice $E_{GV} = |\Omega_{GV}|$ for the condition number considered in Theorem 5.11. We adopt the shorter notation $\mathrm{cond}(\lambda; \Omega_{GV}) \equiv \mathrm{cond}_{|\Omega_{GV}|}(\lambda; \Omega_{GV})$, and we get Theorem 5.12 as a corollary of Theorem 5.11.

**Theorem 5.12.** *Let $C \in \mathbb{R}^{n\times n}$ be a $\{1,1\}$-quasiseparable matrix, let $\lambda \neq 0$ be a simple eigenvalue of $C$ with right eigenvector $\boldsymbol{x}$ and left eigenvector $\boldsymbol{y}$, and let $C = C_L + C_D + C_U$, with $C_L$ strictly lower triangular, $C_D$ diagonal, and $C_U$ strictly upper triangular. Then*

$$\begin{aligned}\mathrm{cond}(\lambda;\Omega_{GV}) = \frac{1}{|\lambda||\boldsymbol{y}^*\boldsymbol{x}|}\Bigg\{ & |\boldsymbol{y}^*||C_D||\boldsymbol{x}| + |\boldsymbol{y}^*C_L||\boldsymbol{x}| + |\boldsymbol{y}^*||C_U\boldsymbol{x}| \\ & + \sum_{i=2}^{n-1}\left|\boldsymbol{y}^*\begin{bmatrix} 0 & 0 \\ -s_i^2\,C(i,1:i-1) & 0 \\ c_i^2\,C(i+1:n,1:i-1) & 0\end{bmatrix}\boldsymbol{x}\right| \\ & + \sum_{j=2}^{n-1}\left|\boldsymbol{y}^*\begin{bmatrix} 0 & -t_j^2C(1:j-1,j) & r_j^2C(1:j-1,j+1:n) \\ 0 & 0 & 0\end{bmatrix}\boldsymbol{x}\right|\Bigg\}.\end{aligned}$$

From the expression in Theorem 5.12 for the relative condition number in the Givens-vector representation for a given $\{1,1\}$-quasiseparable matrix $C \in \mathbb{R}^{n\times n}$, we see that it does not only depend on the matrix entries, the eigenvalue $\lambda$ and on the eigenvectors $\boldsymbol{x}$ and $\boldsymbol{y}$, but it does also depend on the parameters $\{c_i, s_i\}$ and $\{r_i, t_i\}$, which are uniquely determined by the entries of $C$. Therefore, we have obtained an important difference with respect to the relative condition number in a quasiseparable representation presented in Theorem 5.2. Since the Givens-vector representation does not change trivially under diagonal similarities, because of these cosines-sines parameters, this condition number is not invariant under diagonal similarities as we can see from Example 5.13.

**Example 5.13.** *Let $C \in \mathbb{R}^{3\times 3}$ be the $\{1,1\}$-quasiseparable matrix generated as in Theorem 2.21 by the set of Givens-vector parameters given by $\{c_2, s_2\} = \{2.3768 \times 10^{-1}, -9.7134\times 10^{-1}\}$, $\{v_1, v_2\} = \{9.8355, -2.9770\}$, $\{d_1, d_2, d_3\} = \{11.437, -5.3162, 9.7257\}$, $\{e_1, e_2\} = \{1.7658, 9.7074\}$, $\{t_2, r_2\} = \{-9.8216 \times 10^{-1}, 1.8806 \times 10^{-1}\}$, and denote by $\Omega_{GV}^{C}$ the respective tangent-Givens-vector representation. Consider the matrix $K = diag(-1,-1,6)$, and denote by $\Omega_{GV}^{KCK^{-1}}$ the tangent-Givens-vector representation of the matrix $KCK^{-1}$. Then, for the simple eigenvalue $\lambda = 14.120$ and the respective left and right eigenvectors $\boldsymbol{y} = [\,0.96472,\ 0.055889,\ -0.25728\,]^T$ and $\boldsymbol{x} = [\,-0.47887,\ 0.34548,\ 0.80705\,]^T$, of $C$, we have:*

$$cond(\lambda, C; \Omega_{GV}^{C}) = 1.1706 \ \neq \ cond(\lambda, KCK^{-1}; \Omega_{GV}^{KCK^{-1}}) = 1.2485.$$

*Note that for computing $cond(\lambda, KCK^{-1}; \Omega_{GV}^{KCK^{-1}})$, we only need the cosine-sine parameters in $\Omega_{GV}^{KCK^{-1}}$, which can be obtained from the respective tangents parameters computed from the entries of $KCK^{-1}$ as in equation (4.3) in the proof of Lemma 4.15.*

Nevertheless, we will prove in Proposition 5.17 that eigenvalue condition numbers with respect to the tangent-Givens-vector representation can not suffer large variations under diagonal similarities.

## 5.4 Fast computation of the eigenvalue condition number in the Givens-vector representation via tangents

In this section we prove that $\text{cond}(\lambda;\Omega_{GV})$ can be computed fast. This fact is stated in Proposition 5.14, which will be proved by providing an algorithm for computing $\text{cond}(\lambda;\Omega_{GV})$ in $\mathcal{O}(n)$ operations.

**Proposition 5.14.** *Let $C \in \mathbb{R}^{n\times n}$ be a $\{1,1\}$-quasiseparable matrix with a simple eigenvalue $\lambda \neq 0$ with left eigenvector $\boldsymbol{y}$ and right eigenvector $\boldsymbol{x}$, and assume that $\lambda, \boldsymbol{x}, \boldsymbol{y}$, and the Givens-vector representation via tangents $\Omega_{GV}$ of $C$ are all known. Then, cond $(\lambda;\Omega_{GV})$ can be computed in $60n-106$ flops.*

*Proof.* As we did in the proof of Proposition 5.10, we will find the cost of calculating each term in the expression for cond $(\lambda;\Omega_{GV})$ in Theorem 5.12.

(a) In the first place, note that the cost of calculating $c_i = \frac{1}{\sqrt{1+l_i^2}}$ is obviously 4 flops, and the cost of computing $s_i = \frac{l_i}{\sqrt{1+l_i^2}}$, is only 1 flop because we have already calculated the denominator $\sqrt{1+l_i^2}$. Since the same holds for $r_i$ and $t_i$, we conclude that the total cost of calculating $\{c_i, s_i\}_{i=2}^{n-1}$ and $\{r_i, t_i\}_{i=2}^{n-1}$ is $10n-20$ flops.

(b) The total cost of computing $\{c_i^2\}_{i=2}^{n-1}$,$\{s_i^2\}_{i=2}^{n-1}$,$\{r_i^2\}_{i=2}^{n-1}$, and $\{t_i^2\}_{i=2}^{n-1}$ is $4n-8$ flops.

(c) The factor $|\lambda \boldsymbol{y}^*\boldsymbol{x}| = \left|\lambda \sum_{i=1}^{n} \overline{y}_i x_i\right|$ can be computed in $2n$ flops.

(d) Since every product $\overline{y}_i x_i$ has already been calculated in (c), the term

$$|\boldsymbol{y}^*|\,|C_D|\,|\boldsymbol{x}| = \sum_{i=1}^{n} |d_i|\,|\overline{y}_i x_i|,$$

can be calculated in $2n-1$ flops.

(e) For computing the products $\boldsymbol{y}^*C_L$, $C_U\boldsymbol{x}$ (that explicitly appear in the expression for cond $(\lambda;\Omega_{GV})$) and $C_L\boldsymbol{x}$, $\boldsymbol{y}^*C_U$ (which will be, respectively, needed in (g) and (h) in this proof) we can proceed as in (d), (e), (c), (f) in the proof of Proposition 5.10, respectively. In these processes, we obtain simultaneously $\boldsymbol{y}^*C_L^{(-2)}$, $C_U^{(+2)}\boldsymbol{x}$, $C_L^{(-2)}\boldsymbol{x}$, and $\boldsymbol{y}^*C_U^{(+2)}$ (which will also be needed for the fast computation of the last two sums in cond $(\lambda;\Omega_{GV})$). The total cost of these computations is $20n-32$ flops.

(f) The products $|\boldsymbol{y}^* C_L|\,|\boldsymbol{x}|$ and $|\boldsymbol{y}^*|\,|C_U\boldsymbol{x}|$, can be calculated at a cost of $2n-3$ flops each.

(g) Denote $\widetilde{\alpha}_i := \boldsymbol{y}^* \begin{bmatrix} 0 & 0 \\ -s_i^2\, C(i,1:i-1) & 0 \\ c_i^2\, C(i+1:n,1:i-1) & 0 \end{bmatrix} \boldsymbol{x}$, and note that

$$\begin{aligned}\widetilde{\alpha}_i &= \boldsymbol{y}^* \begin{bmatrix} 0 & 0 \\ -s_i^2\, C(i,1:i-1) & 0 \\ 0 & 0 \end{bmatrix} \boldsymbol{x} + \boldsymbol{y}^* \begin{bmatrix} 0 & 0 \\ 0 & 0 \\ c_i^2\, C(i+1:n,1:i-1) & 0 \end{bmatrix} \boldsymbol{x} \\ &= -s_i^2 \overline{y}_i (C_L\boldsymbol{x})_i + c_i^2 \left( \boldsymbol{y}^* \begin{bmatrix} 0 & 0 \\ C(i+1:n,1:i-1) & 0 \end{bmatrix} \boldsymbol{x} \right).\end{aligned}$$

The expression $-s_i^2 \overline{y}_i (C_L\boldsymbol{x})_i$ can be calculated in 2 flops since $s_i^2$ and $C_L\boldsymbol{x}$ have been calculated already. On the other hand, all the expressions

$$\left( \boldsymbol{y}^* \begin{bmatrix} 0 & 0 \\ C(i+1:n,1:i-1) & 0 \end{bmatrix} \boldsymbol{x} \right)$$

can be calculated via a recurrence relation as in (g) in the proof of Proposition 5.10, in $4n-11$ flops. Consequently, the vector $\{\widetilde{\alpha}_i\}_{i=2}^{n-1}$ can be calculated in $8n-19$ flops, and the cost of computing $\sum_{i=2}^{n-1} |\widetilde{\alpha}_i|$ is $9n-22$ flops.

(h) If we denote

$$\widetilde{\beta}_j := \boldsymbol{y}^* \begin{bmatrix} 0 & -t_j^2\, C(1:j-1,j) & r_j^2\, C(1:j-1,j+1:n) \\ 0 & 0 & 0 \end{bmatrix} \boldsymbol{x},$$

then we can proceed in an analogous way to that in (g), and obtain also a cost for computing $\sum_{j=2}^{n-1} \left|\widetilde{\beta}_j\right|$ of $9n-22$ flops.

Finally, by summing all the costs obtained above, and from the expression for $\mathrm{cond}(\lambda;\Omega_{GV})$ we obtain a total cost of $60n-106$ flops for computing $\mathrm{cond}(\lambda;\Omega_{GV})$. □

### 5.4.1 Pseudocode for computing cond $(\lambda;\Omega_{GV})$ fast

In this section we present the pseudocode in Algorithm 2 for computing cond $(\lambda;\Omega_{GV})$ fast. Again, if the reader is not interested in such technical details then this section may be omitted in a first reading.

Since any set of Givens-vector parameters can also be considered as a set of quasiseparable parameters, and since from the proof of Proposition 5.14 we know that cond $(\lambda;\Omega_{GV})$ can be calculated in a very similar way to cond $(\lambda;\Omega_{QS})$, we have that Algorithms 1 and 2 are also similar. Therefore we do not describe Algorithm 2

in detail. We will use the standard functions 'zeros', 'sum', 'conj', and 'ones' from MATLAB. The vectors $\boldsymbol{d}$, zlx, wlx, zly, wly, zux, wux, zuy, wuy, $\alpha$, $\beta$, and the temporary variables tlx1, tly1, tux1, tuy1, tlx2, tly2, tux2, tuy2, are all defined as in Algorithm 1. In the following table, we briefly describe the new variables that appear in the pseudocode of Algorithm 2.

| Variable | Description |
|---|---|
| $\widetilde{\alpha}$ | vector such that $\widetilde{\alpha}(i) = \widetilde{\alpha}_i$ with $\widetilde{\alpha}_i$ as in the proof of Proposition 5.14 |
| $\widetilde{\beta}$ | vector such that $\widetilde{\beta}(i) = \widetilde{\beta}_i$ with $\widetilde{\beta}_i$ as in the proof of Proposition 5.14 |

**Algorithm 2** Fast computation of the eigenvalue condition number $\mathrm{cond}\,(\lambda;\Omega_{GV})$

**Input:** Givens-vector parameters $\{l_i\}_{i=2}^{n-1}, \{v_i\}_{i=1}^{n-1}, \{d_i\}_{i=1}^{n}, \{e_i\}_{i=1}^{n-1}, \{u_i\}_{i=2}^{n-1}$, the eigenvalue $\lambda$ of $C$, the respective left and right eigenvectors $\boldsymbol{y}$ and $\boldsymbol{x}$.

**Set** $c = [\mathrm{ones}(n-2,1)./(\mathrm{sqrt}(\mathrm{ones}(n-2,1) + l.^2))]$, $s = l. * c$, $c = [c;1]$;
$r = [\mathrm{ones}(n-2,1)./(\mathrm{sqrt}(\mathrm{ones}(n-2,1) + u.^2))]$, $t = u. * r$, $r = [r;1]$;
$\mathrm{zlx} = \mathrm{zeros}\,(n,1)$, $\mathrm{wlx} = \mathrm{zeros}\,(n,1)$, $\mathrm{zly} = \mathrm{zeros}\,(1,n)$, $\mathrm{wly} = \mathrm{zeros}\,(1,n)$,
$\mathrm{zux} = \mathrm{zeros}\,(n,1)$, $\mathrm{wux} = \mathrm{zeros}\,(n,1)$, $\mathrm{zuy} = \mathrm{zeros}\,(1,n)$, $\mathrm{wuy} = \mathrm{zeros}\,(1,n)$;
$\mathrm{tlx1} = v_1 \cdot x_1$, $\mathrm{zlx}_2 = c_2 \cdot \mathrm{tlx1}$, $\mathrm{tly1} = \overline{y}_n \cdot c_n$, $\mathrm{zly}_{n-1} = v_{n-1} \cdot \mathrm{tly1}$,
$\mathrm{tux1} = x_n \cdot r_n$, $\mathrm{zux}_{n-1} = e_{n-1} \cdot \mathrm{tux1}$, $\mathrm{tuy1} = e_1 \cdot \overline{y}_1$, $\mathrm{zuy}_2 = r_2 \cdot \mathrm{tuy1}$.
**for** i=3 **to** n **do**
$\mathrm{tlx2} = s_{i-1} \cdot \mathrm{tlx1}$, $\mathrm{wlx}_i = c_i \cdot \mathrm{tlx2}$, $\mathrm{tlx1} = \mathrm{tlx2} + v_{i-1} \cdot x_{i-1}$, $\mathrm{zlx}_i = c_i \cdot \mathrm{tlx1}$;
$\mathrm{tly2} = s_{n-i+1} \cdot \mathrm{tly1}$, $\mathrm{wly}_{n-i+1} = v_{n-i+1} \cdot \mathrm{tly2}$, $\mathrm{tly1} = \mathrm{tly2} + c_{n-i+2} \cdot \overline{y}_{n-i+2}$,
$\mathrm{zly}_{n-i+1} = v_{n-i+1} \cdot \mathrm{tly1}$;
$\mathrm{tux2} = t_{n-i+1} \cdot \mathrm{tux1}$, $\mathrm{wux}_{n-i+1} = e_{n-i+1} \cdot \mathrm{tux2}$, $\mathrm{tux1} = \mathrm{tux2} + r_{n-i+2} \cdot x_{n-i+2}$,
$\mathrm{zux}_{n-i+1} = e_{n-i+1} \cdot \mathrm{tux1}$;
$\mathrm{tuy2} = t_{i-1} \cdot \mathrm{tuy1}$, $\mathrm{wuy}_i = r_i \cdot \mathrm{tuy2}$, $\mathrm{tuy1} = \mathrm{tuy2} + e_{i-1} \cdot \overline{y}_{i-1}$, $\mathrm{zuy}_i = r_i \cdot \mathrm{tuy1}$.
**end for**
**Set** $\alpha = \mathrm{zeros}\,(1,n)$, $\beta = \mathrm{zeros}\,(1,n)$, $\alpha_{n-1} = \overline{y}_n \cdot \mathrm{wlx}_n$, $\beta_2 = \overline{y}_1 \cdot \mathrm{wux}_1$.
$\widetilde{\alpha} = \mathrm{zeros}\,(1,n)$, $\widetilde{\beta} = \mathrm{zeros}\,(1,n)$, $\widetilde{\alpha}_{n-1} = -s_{n-1}^2 \cdot \overline{y}_{n-1} \cdot \mathrm{zlx}_{n-1}$, $\widetilde{\beta}_2 = -t_2^2 \cdot \mathrm{zuy}_2 \cdot x_2$.

**for** i=3 **to** n-1 **do**
$\alpha_{n-i+1} = \alpha_{n-i+2} + \overline{y}_{n-i+2} \cdot \mathrm{wlx}_{n-i+2} - \mathrm{wly}_{n-i+1} \cdot x_{n-i+1}$,
$\widetilde{\alpha}_{n-i+1} = -s_{n-i+1}^2 \cdot \overline{y}_{n-i+1} \cdot \mathrm{zlx}_{n-i+1} + c_{n-i+1}^2 \cdot \alpha_{n-i+1}$,
$\beta_i = \beta_{i-1} + \overline{y}_{i-1} \cdot \mathrm{wux}_{i-1} - \mathrm{wuy}_i \cdot x_i$,
$\widetilde{\beta}_i = -t_i^2 \cdot \mathrm{zuy}_i \cdot x_i + r_i^2 \cdot \beta_i$.
**end for**
**Set** $\mathrm{yx} = \mathrm{conj}(\boldsymbol{y}). * \boldsymbol{x}$;
$\mathrm{cond}\,(\lambda;\Omega_{GV}) = \big(|\boldsymbol{d}'| \cdot |\mathrm{yx}| + |\mathrm{zly}| \cdot |\boldsymbol{x}| + |\boldsymbol{y}'| \cdot |\mathrm{zux}| + \mathrm{sum}\,(|\widetilde{\alpha}|) + \mathrm{sum}(\left|\widetilde{\beta}\right|)\big)/(|\lambda \cdot|\,|\mathrm{sum}(\mathrm{yx})|)$.

**Output:** $\mathrm{cond}\,(\lambda;\Omega_{GV})$.

## 5.5 Comparison of the condition numbers in the quasiseparable and the Givens-vector representation

As we know, the Givens-vector representation is a particular case of the quasiseparable representation which imposes additional constraints on the parameters. Since we will only consider perturbations respecting such constraints, that is, preserving the cosine-sine relations in the parameters $\{c_i, s_i\}$ and $\{r_i, t_i\}$ of $\Omega_{QS}^{GV}$, it is natural to expect $\text{cond}(\lambda; \Omega_{GV})$ not to be larger than $\text{cond}(\lambda; \Omega_{QS})$. In Theorem 5.15, we prove that the Givens-vector representation via tangents is a "more stable" representation than the quasiseparable representation for eigenvalue computations for $\{1,1\}$-quasiseparable matrices. This is the first time that a rigorous proof is given in such direction.

**Theorem 5.15.** *Let $C \in \mathbb{R}^{n\times n}$ be a $\{1,1\}$-quasiseparable matrix, $\Omega_{GV}$ be the tangent-Givens-vector parameters of $C$, $\Omega_{QS}$ be any vector of quasiseparable parameters of $C$, and $\lambda \neq 0$ be a simple eigenvalue of $C$. Then,*

$$\text{cond}(\lambda; \Omega_{GV}) \leq \text{cond}(\lambda; \Omega_{QS}).$$

*Proof.* We compare for the same matrix $C$, the expression given for $\text{cond}(\lambda; \Omega_{QS})$ in Theorem 5.2 and the expression given in Theorem 5.12 for $\text{cond}(\lambda; \Omega_{GV})$. Starting from the sums in the last two terms of the expression for $\text{cond}(\lambda; \Omega_{GV})$, we have:

$$\begin{aligned} S_1 &= \sum_{i=2}^{n-1} \left| \boldsymbol{y}^* \begin{bmatrix} 0 & 0 \\ -s_i^2\, C(i, 1:i-1) & 0 \\ c_i^2\, C(i+1:n, 1:i-1) & 0 \end{bmatrix} \boldsymbol{x} \right| \\ &\leq \sum_{i=2}^{n-1} |\overline{y}_i C_L(i,:)\boldsymbol{x}| + \sum_{i=2}^{n-1} \left| \boldsymbol{y}^* \begin{bmatrix} 0 & 0 \\ C(i+1:n, 1:i-1) & 0 \end{bmatrix} \boldsymbol{x} \right| \\ &= |\boldsymbol{y}^*|\, |C_L \boldsymbol{x}| + \sum_{i=2}^{n-1} \left| \boldsymbol{y}^* \begin{bmatrix} 0 & 0 \\ C(i+1:n, 1:i-1) & 0 \end{bmatrix} \boldsymbol{x} \right|. \end{aligned} \tag{5.4}$$

Analogously, we obtain:

$$\begin{aligned} S_2 &= \sum_{j=2}^{n-1} \left| \boldsymbol{y}^* \begin{bmatrix} 0 & -t_j^2\, C(1:j-1, j) & r_j^2\, C(1:j-1, j+1:n) \\ 0 & 0 & 0 \end{bmatrix} \boldsymbol{x} \right| \\ &\leq |\boldsymbol{y}^* C_U|\, |\boldsymbol{x}| + \sum_{j=2}^{n-1} \left| \boldsymbol{y}^* \begin{bmatrix} 0 & C(1:j-1, j+1:n) \\ 0 & 0 \end{bmatrix} \boldsymbol{x} \right|. \end{aligned} \tag{5.5}$$

From (5.4) and (5.5) we have

$$\text{cond}(\lambda; \Omega_{GV}) \leq \frac{1}{|\lambda||\boldsymbol{y}^*\boldsymbol{x}|} \Big\{ |\boldsymbol{y}^*||C_D||\boldsymbol{x}| + |\boldsymbol{y}^*||C_L\boldsymbol{x}| + |\boldsymbol{y}^* C_L||\boldsymbol{x}|$$

$$
\begin{aligned}
&+ \quad |\boldsymbol{y}^*||C_U\boldsymbol{x}| + |\boldsymbol{y}^*C_U||\boldsymbol{x}| \\
&+ \quad \sum_{i=2}^{n-1} \left| \boldsymbol{y}^* \begin{bmatrix} 0 & 0 \\ C(i+1:n, 1:i-1) & 0 \end{bmatrix} \boldsymbol{x} \right| \\
&+ \quad \sum_{j=2}^{n-1} \left| \boldsymbol{y}^* \begin{bmatrix} 0 & C(1:j-1, j+1:n) \\ 0 & 0 \end{bmatrix} \boldsymbol{x} \right| \Bigg\} \\
&= \quad \mathrm{cond}(\lambda; \Omega_{QS}).
\end{aligned}
$$

□

On the other hand, we are going to show now that the Givens-vector representation via tangents can only improve, with respect to a general quasiseparable representation, the relative condition number of a simple eigenvalue of a quasiseparable matrix, up to a factor of $3n$. Therefore, both representations can be considered equivalent from the point of view of the accuracy of the eigenvalue computations that they allow. This is proved in Theorem 5.16.

**Theorem 5.16.** *Let $\lambda \neq 0$ be a simple eigenvalue of a $\{1,1\}$-quasiseparable matrix $C \in \mathbb{R}^{n\times n}$ and $\Omega_{GV}$ be the tangent-Givens-vector representation of $C$. Then for any quasiseparable representation $\Omega_{QS}$ of $C$:*

$$\frac{\mathrm{cond}(\lambda; \Omega_{QS})}{\mathrm{cond}(\lambda; \Omega_{GV})} \leq 3(n-2).$$

*Proof.* The proof of this theorem is based on Definition 3.20 instead of on the explicit expressions for $\mathrm{cond}(\lambda; \Omega_{QS})$ and $\mathrm{cond}(\lambda; \Omega_{GV})$. Recall that from the Givens-vector representation via tangents $\Omega_{GV}$ of $C$ we can obtain the Givens-vector representation $\Omega_{QS}^{GV}$ of $C$ as in Definition 4.8, and that $\Omega_{QS}^{GV}$ is also a quasiseparable representation of $C$ as we explained after Theorem 2.21. Therefore, in order to use the definition of the componentwise relative eigenvalue condition number for representations, i.e., Definition 3.20, let us consider a quasiseparable perturbation $\delta\Omega_{QS}^{GV}$ of the parameters in $\Omega_{QS}^{GV}$ such that $|\delta\Omega_{QS}^{GV}| \leq \eta|\Omega_{QS}^{GV}|$, and the resulting quasiseparable matrix $\tilde{C} := C(\Omega_{QS}^{GV} + \delta\Omega_{QS}^{GV})$. We will refer to $\eta$ as the *level* of the relative perturbation of the parameters in the representation $\Omega_{QS}^{GV}$. We emphasize that the perturbation $\delta\Omega_{QS}^{GV}$ does not respect in general the pairs cosine-sine.

On the other hand, note that $\tilde{C}$ can also be represented by a vector

$$\Omega'_{GV} := \left(\{l'_i\}_{i=2}^{n-1}, \{v'_i\}_{i=1}^{n-1}, \{d'_i\}_{i=1}^{n}, \{e'_i\}_{i=1}^{n-1}, \{u'_i\}_{i=2}^{n-1}\right)$$

of tangent-Givens-vector parameters and let us consider the perturbations

$$\delta'\Omega_{GV} := \Omega'_{GV} - \Omega_{GV}.$$

For simplicity, we will denote

$$\tilde{C} := C(\Omega_{QS}^{GV} + \delta\Omega_{QS}^{GV}) = C(\Omega'_{GV}).$$

In Lemma 4.15, we provided the upper bound in (5.6) for the level $\eta'$ of the respective relative perturbations in the parameters in $\Omega'_{GV}$ produced by the level $\eta$ of relative perturbation in the quasiseparable parameters in $\Omega_{QS}^{GV}$. More precisely, for a $\{1,1\}$-quasiseparable matrix $C := C(\Omega_{GV}) = C(\Omega_{QS}^{GV})$, it is proved that given any quasiseparable perturbation $|\delta\Omega_{QS}^{GV}| \leq \eta|\Omega_{QS}^{GV}|$, there exists a perturbation $\delta'\Omega_{GV}$ of the tangent-Givens-vector parameters of $C$, such that

$$C(\Omega_{QS}^{GV} + \delta\Omega_{QS}^{GV}) = C(\Omega_{GV} + \delta'\Omega_{GV})$$

and

$$|\delta'\Omega_{GV}| \leq (3(n-2)\eta + \mathcal{O}(\eta^2))|\Omega_{GV}|. \tag{5.6}$$

Then, from Definition 3.20, we have:

$$\operatorname{cond}(\lambda;\Omega_{QS}) \leq \lim_{\eta\to 0}\sup\left\{\frac{|\delta\lambda|}{\eta|\lambda|} : (\lambda+\delta\lambda) \text{ is an eigenvalue of } C(\Omega_{GV}+\delta\Omega_{GV}),\right.$$
$$\left.|\delta\Omega_{GV}| \leq (3(n-2)\eta + \mathcal{O}(\eta^2))|\Omega_{GV}|\right\}.$$

By considering the change of variable $\eta' = (3(n-2)\eta + \mathcal{O}(\eta^2))$, we obtain

$$\operatorname{cond}(\lambda;\Omega_{QS}) \leq \lim_{\eta'\to 0}\sup\left\{\frac{3(n-2)|\delta\lambda|}{\eta'|\lambda|} : (\lambda+\delta\lambda) \text{ is an eigenvalue of } C(\Omega_{GV}+\delta\Omega_{GV}),\right.$$
$$\left.|\delta\Omega_{GV}| \leq \eta'|\Omega_{GV}|\right\} = 3(n-2)\operatorname{cond}(\lambda;\Omega_{GV}).$$

□

From Example 5.13 in Section 5.3 we know that the eigenvalue condition number with respect to the tangent-Givens-vector representation is not invariant under diagonal similarities. However, the variations that may be obtained in the condition number under these similarities are not significant from a numerical point of view. This is stated in Proposition 5.17.

**Proposition 5.17.** *Let $\lambda \neq 0$ be a simple eigenvalue of a $\{1,1\}$-quasiseparable matrix $C \in \mathbb{R}^{n\times n}$, $\Omega_{GV}$ be the tangent-Givens-vector representation of $C$, $K \in \mathbb{R}^{n\times n}$ be diagonal and nonsingular and $\Omega_{GV}^{KCK^{-1}}$ be the tangent-Givens-vector representation of $KCK^{-1}$. Then*

$$\frac{1}{3(n-2)} \leq \frac{\operatorname{cond}(\lambda, C;\Omega_{GV})}{\operatorname{cond}\left(\lambda, KCK^{-1};\Omega_{GV}^{KCK^{-1}}\right)} \leq 3(n-2).$$

*Proof.* The proof is straightforward from Theorems 5.15, 5.16, and 5.9. □

## 5.6 Numerical experiments

In this section, we will discuss briefly some numerical experiments that have been performed in order to confirm some of the results for eigenvalue condition numbers obtained in Sections 5.1, 5.3, and 5.5. We have run several random numerical tests in MATLAB for comparing the unstructured componentwise condition number $\mathrm{cond}(\lambda)$ in Theorem 3.18 and the structured ones $\mathrm{cond}(\lambda;\Omega_{QS})$ and $\mathrm{cond}(\lambda;\Omega_{GV})$. We have started by generating the vectors

$$\boldsymbol{l}\in\mathbb{R}^{n-2},\boldsymbol{v}\in\mathbb{R}^{n-1},\boldsymbol{d}\in\mathbb{R}^{n},\boldsymbol{e}\in\mathbb{R}^{n-1},\text{ and }\boldsymbol{u}\in\mathbb{R}^{n-2}, \tag{5.7}$$

containing the randomly generated parameters of the tangent-Givens-vector representation in Definition 4.8, by using the command `randn` from MATLAB. In addition, in some tests we have scaled the parameters in $\boldsymbol{v}$ and $\boldsymbol{e}$ as explained in (5.8). Then, we build the quasiseparable matrix $C$ of size $n\times n$ generated by these parameters and we obtain its eigenvalues and eigenvectors using the standard command `eig` from MATLAB. The parameters in $\Omega_{QS}:=\Omega_{QS}^{GV}$ in Definition 4.8 are also computed. Finally, we compute the structured eigenvalue condition numbers $\mathrm{cond}(\lambda;\Omega_{QS})$ and $\mathrm{cond}(\lambda;\Omega_{GV})$, using our fast Algorithms 1 and 2 respectively, and the unstructured condition number $\mathrm{cond}(\lambda)$ using direct matrix-vector multiplication with a resulting cost of order $2n^2+\mathcal{O}(n)$ operations (note that $\mathrm{cond}(\lambda)$ can also be computed in $\mathcal{O}(n)$ operations since $|C|$ is also a quasiseparable matrix), and we compare these three condition numbers.

When the generated parameters are completely random (i.e., no scaling has been introduced), we have not observed large differences between the eigenvalue condition numbers, i.e., $\mathrm{cond}(\lambda)\approx\mathrm{cond}(\lambda;\Omega_{QS})\approx\mathrm{cond}(\lambda;\Omega_{GV})$, and all of them are very often moderate.

On the other hand, as announced above, we have also performed tests where some scalings over the randomly generated parameters of the tangent-Givens-vector representation have been introduced in order to build an *unbalanced* quasiseparable matrix. After generating the random vectors as in (5.7), we have scaled the parameters in $\boldsymbol{v}$ and $\boldsymbol{e}$ by creating the vectors (using MATLAB standard notation):

$$\boldsymbol{scv}=(10.\hat{}(k:-(k-1)/(n-2):1))\text{ and }\boldsymbol{sce}=(10.\hat{}(1:(k-1)/(n-2):k))\,,$$

where $k$ is a fixed natural number not greater than 10 in our experiments, and considering

$$\boldsymbol{v}=10^2*\boldsymbol{scv}.*\boldsymbol{v}\text{ and }\boldsymbol{e}=10^2*\boldsymbol{sce}.*\boldsymbol{e}\,, \tag{5.8}$$

which may produce unbalanced rows and columns in the lower and upper triangular parts of the matrix generated by these parameters (see Example 2.22). Therefore, we may expect large unstructured eigenvalue condition numbers. In this way, for different values of $k$ and different sizes $n$, we have found distributions of the tangent-Givens-vector parameters that produce $\{1,1\}$-quasiseparable matrices such that the

unstructured condition number $\text{cond}(\lambda)$ is much larger than the structured ones, $\text{cond}(\lambda;\Omega_{QS})$ and $\text{cond}(\lambda;\Omega_{GV})$. In fact, for $n = 200$ and $n = 300$, with $k = 5$ in both cases, we have obtained particular matrices such that:

| $n$ | $\lambda_{\min}$ | $\lambda_{\max}$ | $\max_{\lambda}\left\{\frac{\text{cond}(\lambda)}{\text{cond}(\lambda;\Omega_{QS})}\right\}$ | $\max_{\lambda}\left\{\frac{\text{cond}(\lambda;\Omega_{QS})}{\text{cond}(\lambda;\Omega_{GV})}\right\}$ |
|---|---|---|---|---|
| 200 | 555.2761 | $-2.4628\cdot 10^4 + 1.6840\cdot 10^5 i$ | $4.7847\cdot 10^{11}$ | 3.3337 |
| 300 | 60.5936 | $-1.3533\cdot 10^5 + 2.6060\cdot 10^5 i$ | $2.4201\cdot 10^{10}$ | 3.2570 |

,

where $\lambda_{\min}$ and $\lambda_{\max}$ denote the respective minimum and maximum eigenvalues in absolute value, while $\max_{\lambda}\left\{\frac{\text{cond}(\lambda)}{\text{cond}(\lambda;\Omega_{QS})}\right\}$ and $\max_{\lambda}\left\{\frac{\text{cond}(\lambda;\Omega_{QS})}{\text{cond}(\lambda;\Omega_{GV})}\right\}$ denote the respective maximum values of the quotients, both taken over the corresponding sets of the eigenvalues of the particular considered matrices for $n = 200$ and $n = 300$. Furthermore, if we denote by $\lambda_{\text{opt}}$ the eigenvalue where the maximum of the quotient $\left\{\frac{\text{cond}(\lambda)}{\text{cond}(\lambda;\Omega_{QS})}\right\}$ occurs, we have:

| $n$ | $\lambda_{\text{opt}}$ | $\text{cond}(\lambda_{\text{opt}})$ | $\text{cond}(\lambda_{\text{opt}};\Omega_{QS})$ |
|---|---|---|---|
| 200 | $-2.4569\cdot 10^3 + 3.7791\cdot 10^2 i$ | $2.4780\cdot 10^{13}$ | 51.7901 |
| 300 | $-1.0088\cdot 10^4$ | $1.8703\cdot 10^{11}$ | 7.7281 |

.

It is worth mentioning that, as one would expect, repeating these experiments for values of $k$ greater than 5 (which produce matrices with strongly unbalanced lower and upper triangular parts), we obtained similar results to the ones above.

These examples show not only that the structured eigenvalue condition number $\text{cond}(\lambda;\Omega_{QS})$ (and, therefore, $\text{cond}(\lambda;\Omega_{GV})$) may be much smaller than the unstructured one $\text{cond}(\lambda)$, but also that there exist $\{1,1\}$-quasiseparable matrices having eigenvalues that are very ill conditioned with respect to perturbations of its entries, but that are very well conditioned with respect to perturbations of its quasiseparable parameters.

In addition, as one would expect from Theorems 5.15 and 5.16, there is not much difference between the values of $\text{cond}(\lambda;\Omega_{QS})$ and $\text{cond}(\lambda;\Omega_{GV})$, therefore we have omitted the corresponding column for the quotient $\frac{\text{cond}(\lambda;\Omega_{QS})}{\text{cond}(\lambda;\Omega_{GV})}$ in the second table above. In fact, in all our tests we have obtained that

$$\frac{\text{cond}(\lambda;\Omega_{QS})}{\text{cond}(\lambda;\Omega_{GV})} < 15,$$

which suggests that the bound in Theorem 5.16 may be improved up to a constant.

# Chapter 6

# Structured eigenvalue condition numbers for $\{n_L, n_U\}$-quasiseparable matrices

In this chapter we will study structured eigenvalue condition numbers for $\{n_L, n_U\}$-quasiseparable matrices in an analogous way as we did for $\{1, 1\}$-quasiseparable matrices in Chapter 5, but with respect to the general quasiseparable representation described in Section 2.3.3. This representation has important differences with respect to the $\{1, 1\}$-case, since its parameters are the entries of certain vectors and matrices. This means that we are in a more complicated scenario, since the interactions (via products) between the parameters that generate the matrix are no longer trivial because they involve matrix and vector multiplications. Consequently, in Section 6.1, we will provide an expression for computing the eigenvalue condition number with respect to the quasiseparable representation of an $\{n_L, n_U\}$-quasiseparable matrix (see Theorem 6.2), and we will note that this structured eigenvalue condition number may be much larger than the usual unstructured one in Definition 3.17, as observed in the numerical experiments described in Section 6.3. This result represents an important difference with respect to the one obtained in Proposition 5.4 for the $\{1, 1\}$-case and suggests the use of a representation different than the quasiseparable one for general $\{n_L, n_U\}$-quasiseparable matrices which can be potentially more stable. Nevertheless, we provide a bound for the structured eigenvalue condition number in the quasiseparable representation in terms of an unstructured eigenvalue condition number defined by Higham and Higham in [42]. In addition, in Section 6.2, we describe how to compute the structured eigenvalue condition number fast.

## 6.1 Eigenvalue condition numbers for $\{n_L, n_U\}$-quasiseparable matrices in the quasiseparable representation

In an analogous way to the quasiseparable representation for $\{1,1\}$-quasiseparable matrices, we can deduce relative eigenvalue condition numbers with respect to componentwise perturbations of the parameters in the quasiseparable representation for $\{n_L, n_U\}$-quasiseparable matrices, although the corresponding expressions are considerably more involved that in the $\{1,1\}$-case. For simplicity, we will only consider relative perturbations with respect to the parameters in the representation. Therefore, we will use Theorem 3.23 for proving Theorem 6.2, where the new condition number is deduced. First, in order to avoid very complicated expressions, we need to define the *Kronecker product*. This is done in Definition 6.1 [44].

**Definition 6.1.** *For any two matrices $A = [a_{ij}] \in \mathbb{C}^{m\times n}$ and $B = [b_{ij}] \in \mathbb{C}^{p\times q}$, we define the* Kronecker product *of $A$ and $B$ (in that order) as the block matrix:*

$$A \otimes B = \begin{bmatrix} a_{11}B & \cdots & a_{1n}B \\ \vdots & \ddots & \vdots \\ a_{m1}B & \cdots & a_{mn}B \end{bmatrix} \in \mathbb{C}^{mp\times nq}.$$

On the other hand, before Theorem 6.2 we need to introduce some notation. For an $\{n_L, n_U\}$-quasiseparable matrix $C = C_L + C_D + C_U$, with $C_L$ strictly lower triangular, $C_D$ diagonal, $C_U$ strictly upper triangular, of size $n \times n$, and with a quasiseparable representation $\Omega_{QS}$ as in Theorem 2.23, we denote:

- $P \in \mathbb{R}^{1\times n_L n}$ the vector partitioned as $P = [0_{1\times n_L}|p_2|\cdots|p_n]$,
- $\mathrm{diag}(P)$ the diagonal matrix of size $n_L n \times n_L n$ that is obtained by placing the elements of $P$ on its diagonal entries, i.e.:

$$\mathrm{diag}(P) = \begin{bmatrix} 0_{n_L\times n_L} & & & & & \\ & (p_2)_1 & & & & \\ & & \ddots & & & \\ & & & (p_2)_{n_L} & & \\ & & & & \ddots & \\ & & & & & (p_n)_{n_L} \end{bmatrix},$$

- $C_L[p]$ the matrix of size $n_L n \times n$, partitioned into $n \times n$ blocks of size $n_L \times 1$, which is obtained by removing the parameters $p_i$ from $C_L$, adding $n_L$ rows of

zeros on the top and 1 column of zeros on the right, i.e.:

$$C_L[p] = \begin{bmatrix} 0_{n_L \times 1} & & & \\ & 0_{n_L \times 1} & & 0 \\ a_{ij}^{\times} q_j & & \ddots & \\ & & & 0_{n_L \times 1} \end{bmatrix},$$

- $Q \in \mathbb{R}^{n_L n \times 1}$ the vector partitioned as

$$Q = \begin{pmatrix} q_1 \\ \vdots \\ q_{n-1} \\ 0_{n_L \times 1} \end{pmatrix},$$

- $\mathrm{diag}(Q)$ the diagonal matrix of size $n_L n \times n_L n$ that is obtained by placing the elements of $Q$ on its diagonal entries, i.e.:

$$\mathrm{diag}(Q) = \begin{bmatrix} (q_1)_1 & & & & & & \\ & \ddots & & & & & \\ & & (q_1)_{n_L} & & & & \\ & & & \ddots & & & \\ & & & & (q_{n-1})_{n_L} & & \\ & & & & & 0_{n_L \times n_L} \end{bmatrix},$$

- $C_L[q]$ the matrix of size $n \times n n_L$, partitioned into $n \times n$ blocks of size $1 \times n_L$, which is obtained by removing the parameters $q_j$ from $C_L$, adding $n_L$ columns of zeros on the right and 1 row of zeros on the top, i.e.:

$$C_L[q] = \begin{bmatrix} 0_{1 \times n_L} & & & \\ & 0_{1 \times n_L} & & 0 \\ p_i a_{ij}^{\times} & & \ddots & \\ & & & 0_{1 \times n_L} \end{bmatrix},$$

- $A \in \mathbb{R}^{n_L n \times n_L n}$ a block diagonal matrix partioned into $n \times n$ blocks of size $n_L \times n_L$, such that the $n_L \times n_L$ diagonal block $A_{ii}$ is given by the matrix $a_i$ for $i$ in $\{2, \ldots, n-1\}$, and such that the $n_L \times n_L$ blocks $A_{11}$ and $A_{nn}$ are both zero matrices, i.e.:

$$A = \begin{bmatrix} 0_{n_L \times n_L} & & & & & \\ & a_2 & & & & \\ & & a_3 & & & \\ & & & \ddots & & \\ & & & & a_{n-1} & \\ & & & & & 0_{n_L \times n_L} \end{bmatrix},$$

- $G \in \mathbb{R}^{1\times n_U n}$ the vector partitioned as $G = [g_1|\cdots|g_{n-1}|0_{1\times n_U}]$,

- $\mathrm{diag}(G)$ a diagonal matrix of size $n_U n \times n_U n$ that is obtained by placing the elements of $G$ on its diagonal entries, i.e.:

$$\mathrm{diag}(G) = \begin{bmatrix} (g_1)_1 & & & & & & \\ & \ddots & & & & & \\ & & (g_1)_{n_U} & & & & \\ & & & \ddots & & & \\ & & & & (g_{n-1})_{n_U} & & \\ & & & & & 0_{n_U\times n_U} \end{bmatrix},$$

- $C_U[g]$ the matrix of size $n_U n \times n$, partitioned into $n \times n$ blocks of size $n_U \times 1$, which is obtained by removing the parameters $g_i$ from $C_U$, adding $n_U$ rows of zeros on the bottom, and one column of zeros on the left, i.e.:

$$C_U[g] = \begin{bmatrix} 0_{n_U\times 1} & & & \\ & 0_{n_U\times 1} & & b_{ij}^{\times} h_j \\ 0 & & \ddots & \\ & & & 0_{n_U\times 1} \end{bmatrix},$$

- $H \in \mathbb{R}^{n_U n\times 1}$ the vector partitioned as follows

$$H = \begin{pmatrix} 0_{n_U\times 1} \\ h_2 \\ \vdots \\ h_n \end{pmatrix},$$

- $\mathrm{diag}(H)$ a diagonal matrix of size $n_U n \times n_U n$ that is obtained by placing the elements of $H$ on its diagonal entries, i.e.:

$$\mathrm{diag}(H) = \begin{bmatrix} 0_{n_U\times n_U} & & & & & \\ & (h_2)_1 & & & & \\ & & \ddots & & & \\ & & & (h_2)_{n_U} & & \\ & & & & \ddots & \\ & & & & & (h_n)_{n_U} \end{bmatrix},$$

- $C_U[h]$ the matrix of size $n \times nn_U$, partitioned into $n \times n$ blocks of size $1 \times n_U$, which is obtained by removing the parameters $h_j$ from $C_U$ and adding $n_U$ columns of zeros on the left and one row of zeros on the bottom, i.e.:

$$C_U[h] = \begin{bmatrix} 0_{1\times n_U} & & & \\ & 0_{1\times n_U} & & g_i b_{ij}^{\times} \\ 0 & & \ddots & \\ & & & 0_{1\times n_U} \end{bmatrix}, \text{ and}$$

- $B \in \mathbb{R}^{n_U n \times n_U n}$ a block diagonal matrix partioned into $n \times n$ blocks of size $n_U \times n_U$, such that the $n_U \times n_U$ diagonal block $B_{ii}$ is given by the matrix $b_i$ for $i$ in $\{2, \ldots, n-1\}$, and such that the $n_U \times n_U$ blocks $B_{11}$ and $B_{nn}$ are both zero matrices, i.e.

$$B = \begin{bmatrix} 0_{n_U\times n_U} & & & & & \\ & b_2 & & & & \\ & & b_3 & & & \\ & & & \ddots & & \\ & & & & b_{n-1} & \\ & & & & & 0_{n_U\times n_U} \end{bmatrix}.$$

**Theorem 6.2.** *Let $C \in \mathbb{R}^{n\times n}$ be an $\{n_L, n_U\}$-quasiseparable matrix and $C = C_L + C_D + C_U$, with $C_L$ strictly lower triangular, $C_D$ diagonal, and $C_U$ strictly upper triangular. Suppose $\lambda \neq 0$ is a simple eigenvalue of $C$ with left and right eigenvectors $\boldsymbol{y}$ and $\boldsymbol{x}$ respectively, and denote by $\Omega_{QS}$ a quasiseparable representation of $C$ as in Theorem 2.23. Then, using the notation above, the componentwise relative condition number* $\mathrm{cond}(\lambda; \Omega_{QS})$ *of $\lambda$ with respect to $\Omega_{QS}$ is given by the following expression:*

$$\begin{aligned} \mathrm{cond}(\lambda; \Omega_{QS}) = \frac{1}{|\lambda||\boldsymbol{y}^*\boldsymbol{x}|} \Big\{ & |\boldsymbol{y}^*||C_D|||\boldsymbol{x}| + |\boldsymbol{y}^* \otimes \boldsymbol{e}_{n_L}^T||\mathrm{diag}(P)||C_L[p]\boldsymbol{x}| \\ & + |\boldsymbol{y}^* C_L[q]||\mathrm{diag}(Q)||\boldsymbol{x} \otimes \boldsymbol{e}_{n_L}| \\ & + |\boldsymbol{y}^* \otimes \boldsymbol{e}_{n_U}^T||\mathrm{diag}(G)||C_U[g]\boldsymbol{x}| \\ & + |\boldsymbol{y}^* C_U[h]||\mathrm{diag}(H)||\boldsymbol{x} \otimes \boldsymbol{e}_{n_U}| \\ & + |\boldsymbol{y}^* C_L[q]|\,|A|\,|C_L[p]\boldsymbol{x}| \\ & + |\boldsymbol{y}^* C_U[h]|\,|B|\,|C_U[g]\boldsymbol{x}| \Big\}, \end{aligned}$$

*where:*
$\boldsymbol{e}_{n_L}^T \in \mathbb{R}^{1\times n_L}$ *denotes the vector* $\boldsymbol{e}_{n_L}^T = [1, 1, \ldots, 1]$ *and* $\boldsymbol{e}_{n_U}^T \in \mathbb{R}^{1\times n_U}$ *denotes the vector* $\boldsymbol{e}_{n_U}^T = [1, 1, \ldots, 1]$.

*Proof.* Using Theorem 3.23 and following the ideas in the proofs of Theorems 5.2 and 5.12, we will proceed by calculating the partial derivatives of the eigenvalue with respect to the different parameters, term by term, as follows.

**Contribution of the derivatives with respect to $\{d_i\}_{i=1}^n$ :**

From

$$\frac{d_i}{\lambda}\frac{\partial \lambda}{\partial d_i} = \frac{1}{\lambda(\boldsymbol{y}^*\boldsymbol{x})}\overline{y}_i d_i x_i,$$

we have

$$\sum_{i=1}^{n}\left|\frac{d_i}{\lambda}\frac{\partial \lambda}{\partial d_i}\right| = \frac{1}{|\lambda(\boldsymbol{y}^*\boldsymbol{x})|}|\boldsymbol{y}^*||C_D||\boldsymbol{x}|.$$

**Contribution of the derivatives with respect to the entries of $\{p_i\}_{i=2}^n$ :**

Recall that $p_i = [(p_i)_1, (p_i)_2, \ldots, (p_i)_{n_L}]$. Then

$$\begin{aligned}\frac{(p_i)_l}{\lambda}\frac{\partial \lambda}{\partial (p_i)_l} &= \frac{(p_i)_l}{\lambda(\boldsymbol{y}^*\boldsymbol{x})}\left\{\boldsymbol{y}^*\begin{bmatrix} 0 & & & & & \\ & 0 & & & & \\ & & \ddots & & & \\ \frac{\partial C_{i1}}{\partial (p_i)_l} & \cdots & \frac{\partial C_{i,i-1}}{\partial (p_i)_l} & 0 & & \\ & & & & \ddots & \\ & & & & & 0 \end{bmatrix}\boldsymbol{x}\right\} \\ &= \frac{(p_i)_l}{\lambda(\boldsymbol{y}^*\boldsymbol{x})}\overline{y}_i\left[\frac{\partial C_{i1}}{\partial (p_i)_l}, \ldots, \frac{\partial C_{i,i-1}}{\partial (p_i)_l}, 0, \ldots, 0\right]\boldsymbol{x},\end{aligned}$$

and

$$\sum_{l=1}^{n_L}\left|\frac{(p_i)_l}{\lambda}\frac{\partial \lambda}{\partial (p_i)_l}\right| = \frac{|y_i|}{|\lambda(\boldsymbol{y}^*\boldsymbol{x})|}\,|p_i|\,\left|[a_{i1}^{\times}q_1|\cdots|a_{i,i-1}^{\times}q_{i-1}|0|\cdots|0]\boldsymbol{x}\right|.$$

Therefore, the global contribution of the derivatives with respect to the parameters in $\{p_i\}_{i=2}^n$ is stated in the following expression:

$$\sum_{i=2}^{n}\sum_{l=1}^{n_L}\left|\frac{(p_i)_l}{\lambda}\frac{\partial \lambda}{\partial (p_i)_l}\right| = \frac{1}{|\lambda(\boldsymbol{y}^*\boldsymbol{x})|}\sum_{i=2}^{n}|y_i|\,|p_i|\,\left|[a_{i1}^{\times}q_1|\cdots|a_{i,i-1}^{\times}q_{i-1}|0|\cdots|0]\boldsymbol{x}\right|,$$

which can be rewritten as:

$$\sum_{i=2}^{n}\sum_{l=1}^{n_L}\left|\frac{(p_i)_l}{\lambda}\frac{\partial \lambda}{\partial (p_i)_l}\right| = \frac{1}{|\lambda(\boldsymbol{y}^*\boldsymbol{x})|}\left|\boldsymbol{y}^*\otimes \boldsymbol{e}_{n_L}^T\right|\,|\mathrm{diag}(P)|\,|C_L[p]\boldsymbol{x}|\,.$$

**Contribution of the derivatives with respect to the entries of $\{a_i\}_{i=2}^{n-1}$ :**

Recall that $a_i = \{(a_i)_{lm}\}_{l,m=1}^{n_L}$ and let us denote $C[(a_i)_{lm}]$ to the matrix of size $n \times n$ that it is obtained from $C$ by replacing the matrix $a_i$ by the matrix $(a_i)_{lm}E_{lm}$, where $E_{lm} \in \mathbb{R}^{n_L \times n_L}$, and $(E_{lm})_{st} = \delta_{ls}\delta_{mt}$, where $\delta_{ls}$ denotes the Kronecker delta, i.e.:

$$\delta_{ls} = \begin{cases} 1 \text{ if } l = s \\ 0 \text{ if } l \neq s \end{cases} .$$

Then, using the notation above, the following equalities are straightforward.

$$\begin{aligned} \frac{(a_i)_{lm}}{\lambda}\frac{\partial \lambda}{\partial (a_i)_{lm}} &= \frac{(a_i)_{lm}}{\lambda(\boldsymbol{y}^*\boldsymbol{x})}\boldsymbol{y}^* \begin{bmatrix} 0 & 0 \\ \frac{\partial C(i+1:n,1:i-1)}{\partial (a_i)_{lm}} & 0 \end{bmatrix} \boldsymbol{x} \\ &= \frac{1}{\lambda(\boldsymbol{y}^*\boldsymbol{x})}\boldsymbol{y}^* \begin{bmatrix} 0 & 0 \\ C[(a_i)_{lm}](i+1:n,1:i-1) & 0 \end{bmatrix} \boldsymbol{x}. \end{aligned}$$

Therefore

$$\sum_{i=2}^{n-1}\sum_{l,m=1}^{n_L}\left|\frac{(a_i)_{lm}}{\lambda}\frac{\partial \lambda}{\partial (a_i)_{lm}}\right| = \frac{1}{|\lambda(\boldsymbol{y}^*\boldsymbol{x})|}\sum_{i=2}^{n-1}\sum_{l,m=1}^{n_L}\left|\boldsymbol{y}^* \begin{bmatrix} 0 & 0 \\ C[(a_i)_{lm}](i+1:n,1:i-1) & 0 \end{bmatrix} \boldsymbol{x}\right|.$$

Let us focus on the matrix in the sum above, and denote:

$$A_{i;lm} = \begin{bmatrix} 0 & 0 \\ C[(a_i)_{lm}](i+1:n,1:i-1) & 0 \end{bmatrix}.$$

Let us rewrite $A_{i;lm}$ in a more explicit way:

$$A_{i;lm} = \begin{bmatrix} 0 & \cdots & 0 & 0 \\ p_{i+1}\left((a_i)_{lm}E_{lm}\right)a_{i-1}\cdots a_2 q_1 & \cdots & p_{i+1}\left((a_i)_{lm}E_{lm}\right)q_{i-1} & 0 \\ \vdots & \ddots & \vdots & \vdots \\ p_{i+k}a_{i+k-1}\cdots a_{i+1}\left((a_i)_{lm}E_{lm}\right)a_{i-1}\cdots a_2 q_1 & \cdots & p_{i+k}a_{i+k-1}\cdots a_{i+1}\left((a_i)_{lm}E_{lm}\right)q_{i-1} & 0 \\ \vdots & \ddots & \vdots & \vdots \\ p_n a_{n-1}\cdots a_{i+1}\left((a_i)_{lm}E_{lm}\right)a_{i-1}\cdots a_2 q_1 & \cdots & p_n a_{n-1}\cdots a_{i+1}\left((a_i)_{lm}E_{lm}\right)q_{i-1} & 0 \end{bmatrix},$$

which can be rewritten again, as the following product:

$$A_{i;lm} = \begin{bmatrix} 0 \\ p_{i+1} \\ \vdots \\ p_{i+k}a_{i+k-1}\ldots a_{i+1} \\ \vdots \\ p_n a_{n-1}\cdots a_{i+1} \end{bmatrix} [(a_i)_{lm}E_{lm}] \left[\begin{array}{c|c|c|c|c} a_{i-1}\ldots a_2 q_1 & \cdots & a_{i-1}\ldots a_{j+1}q_j & \cdots & q_{i-1} \end{array}\middle|\, 0 \right].$$

Note now that from the respective definitions of the matrices $C_L[q]$ and $C_L[p]$, we have:

$$C_L[q](:,i) = \begin{bmatrix} 0 \\ p_{i+1} \\ \vdots \\ p_{i+k}a_{i+k-1}\dots a_{i+1} \\ \vdots \\ p_n a_{n-1}\cdots a_{i+1} \end{bmatrix}$$

and

$$C_L[p](i,:) = \left[\; a_{i-1}\dots a_2 q_1 \;\middle|\; \cdots \;\middle|\; a_{i-1}\dots a_{j+1}q_j \;\middle|\; \cdots \;\middle|\; q_{i-1} \;\middle|\; 0 \;\right],$$

from where we obtain

$$A_{i;lm} = C_L[q](:,i)\,[(a_i)_{lm}E_{lm}]\,C_L[p](i,:).$$

Then, since no additions occur in the multiplication by $[(a_i)_{lm}E_{lm}]$ in the equation above, it is easy to see that

$$\begin{aligned}\sum_{l,m=1}^{n_L} |\boldsymbol{y}^* A_{i;lm}\boldsymbol{x}| &= \sum_{l,m=1}^{n_L} |\boldsymbol{y}^* C_L[q](:,i)\,[(a_i)_{lm}E_{lm}]\,C_L[p](i,:)\boldsymbol{x}| \\ &= |\boldsymbol{y}^* C_L[q](:,i)|\,|a_i|\,|C_L[p](i,:)\boldsymbol{x}|\,;\end{aligned}$$

and we conclude that

$$\begin{aligned}\sum_{i=2}^{n-1}\sum_{l,m=1}^{n_L}\left|\frac{(a_i)_{lm}}{\lambda}\frac{\partial\lambda}{\partial(a_i)_{lm}}\right| &= \frac{1}{|\lambda(\boldsymbol{y}^*\boldsymbol{x})|}\sum_{i=2}^{n-1}|\boldsymbol{y}^* C_L[q](:,i)|\,|a_i|\,|C_L[p](i,:)\boldsymbol{x}| \\ &= \frac{1}{|\lambda(\boldsymbol{y}^*\boldsymbol{x})|}\,|\boldsymbol{y}^* C_L[q]|\,|A|\,|C_L[p]\boldsymbol{x}|\,.\end{aligned}$$

**Contribution of the derivatives with respect to the entries of $\{q_j\}_{j=1}^{n-1}$ :**

Recall that $q_j = [(q_j)_1, \dots, (q_j)_{n_L}]^T$ . Then,

$$\frac{(q_j)_l}{\lambda}\frac{\partial\lambda}{\partial(q_j)_l} = \frac{(q_j)_l}{\lambda(\boldsymbol{y}^*\boldsymbol{x})}\left(\boldsymbol{y}^*\begin{bmatrix} 0 & & & & \\ & \ddots & & & \\ & & 0 & & \\ & & \frac{\partial C_{j+1,j}}{\partial(q_j)_l} & 0 & \\ & & \vdots & & \ddots \\ & & \frac{\partial C_{n,j}}{\partial(q_j)_l} & & 0 \end{bmatrix}\boldsymbol{x}\right)$$

$$= \frac{(q_j)_l}{\lambda(\boldsymbol{y}^*\boldsymbol{x})}\boldsymbol{y}^* \begin{bmatrix} 0 \\ \vdots \\ 0 \\ \frac{\partial C_{j+1,j}}{\partial (q_j)_l} \\ \vdots \\ \frac{\partial C_{n,j}}{\partial (q_j)_l} \end{bmatrix} x_j.$$

Therefore, the global contribution of the derivatives with respect to the parameters in $\{q_j\}_{j=1}^{n-1}$ is stated in the following expression.

$$\sum_{j=1}^{n-1}\sum_{l=1}^{n_L}\left|\frac{(q_j)_l}{\lambda}\frac{\partial\lambda}{\partial(q_j)_l}\right| = \frac{1}{|\lambda(\boldsymbol{y}^*\boldsymbol{x})|}\sum_{j=1}^{n-1}\left|\boldsymbol{y}^* \begin{bmatrix} 0 \\ \vdots \\ 0 \\ p_{j+1}a^{\times}_{j+1,j} \\ \vdots \\ p_n a^{\times}_{n,j} \end{bmatrix}\right| |(q_j)|\,|x_j|,$$

which can be rewritten as:

$$\sum_{j=1}^{n-1}\sum_{l=1}^{n_L}\left|\frac{(q_j)_l}{\lambda}\frac{\partial\lambda}{\partial(q_j)_l}\right| = \frac{1}{|\lambda(\boldsymbol{y}^*\boldsymbol{x})|}\,|\boldsymbol{y}^*C_L[q]|\,|\mathrm{diag}(Q)|\,|\boldsymbol{x}\otimes\boldsymbol{e}_{n_L}|.$$

By proceeding in an analogous way to what it has been done above for the parameters in $\{p_i\}_{i=2}^{n}, \{a_i\}_{i=2}^{n-1}$, and $\{q_j\}_{j=1}^{n-1}$ representing the strictly lower triangular part of $C$, we can obtain the corresponding contribution to $\mathrm{cond}(\lambda;\Omega_{QS})$ of the parameters in $\{g_i\}_{i=1}^{n-1}, \{b_i\}_{i=2}^{n-1}$, and $\{h_j\}_{j=2}^{n}$ representing the strictly upper triangular part of $C$.

**Contribution of the derivatives with respect to the entries of $\{g_i\}_{i=1}^{n-1}$ :**

$$\sum_{i=1}^{n-1}\sum_{l=1}^{n_U}\left|\frac{(g_i)_l}{\lambda}\frac{\partial\lambda}{\partial(g_i)_l}\right| = \frac{1}{|\lambda(\boldsymbol{y}^*\boldsymbol{x})|}\left|\boldsymbol{y}^*\otimes\boldsymbol{e}_{n_U}^T\right|\,|\mathrm{diag}(G)|\,|C_U[g]\boldsymbol{x}|\,.$$

**Contribution of the derivatives with respect to the entries of $\{b_i\}_{i=2}^{n-1}$ :**

Denote $C[(b_i)_{lm}]$ to the matrix of size $n\times n$ that it is obtained from $C$ by replacing the matrix $b_i$ by the matrix $(b_i)_{lm}E_{lm}$, where $E_{lm}\in\mathbb{R}^{n_U\times n_U}$, and $(E_{lm})_{st}=\delta_{ls}\delta_{mt}$. Then,

$$\sum_{i=2}^{n-1}\sum_{l,m=1}^{n_U}\left|\frac{(b_i)_{lm}}{\lambda}\frac{\partial\lambda}{\partial(b_i)_{lm}}\right| = \frac{1}{|(\lambda\boldsymbol{y}^*\boldsymbol{x})|}\sum_{i=2}^{n-1}\sum_{l,m=1}^{n_U}\left|\boldsymbol{y}^*\begin{bmatrix} 0 & C[(b_i)_{lm}](1:i-1,i+1:n) \\ 0 & 0 \end{bmatrix}\boldsymbol{x}\right|$$

$$= \frac{1}{|\lambda(\boldsymbol{y}^*\boldsymbol{x})|} |\boldsymbol{y}^* C_U[h]| \, |B| \, |C_U[g]\boldsymbol{x}| \, .$$

**Contribution of the derivatives with respect to the entries of $\{h_j\}_{j=2}^n$ :**

$$\sum_{j=2}^{n} \sum_{l=1}^{n_U} \left| \frac{(h_j)_l}{\lambda} \frac{\partial \lambda}{\partial (h_j)_l} \right| = \frac{1}{|\lambda(\boldsymbol{y}^*\boldsymbol{x})|} |\boldsymbol{y}^* C_U[h]| \, |\mathrm{diag}(H)| \, |\boldsymbol{x} \otimes \boldsymbol{e}_{n_U}|.$$

The proof is completed by summing all the global contributions obtained above for the derivatives of the parameters in $\Omega_{QS}$. □

**Remark 6.3.** *Note that Theorem 6.2, stated for $\{n_L, n_U\}$-quasiseparable matrices, is a generalization of Theorem 5.2, stated for $\{1,1\}$-quasiseprable matrices . In fact, for $n_L = n_U = 1$, the following equalities are straightforward,*

$$\begin{array}{l} \boldsymbol{y}^* \otimes \boldsymbol{e}_{n_L}^T = \boldsymbol{y}^*, \; \boldsymbol{x} \otimes \boldsymbol{e}_{n_U} = \boldsymbol{x}, \; \boldsymbol{y}^* \otimes \boldsymbol{e}_{n_U}^T = \boldsymbol{y}^*, \; \boldsymbol{x} \otimes \boldsymbol{e}_{n_U} = \boldsymbol{x}, \\ |diag(P)| \, |C_L[p]\boldsymbol{x}| = |C_L\boldsymbol{x}| \, , \; |\boldsymbol{y}^* C_L[q]| \, |diag(Q)| = |\boldsymbol{y}^* C_L| \, , \\ |diag(G)| \, |C_U[g]\boldsymbol{x}| = |C_U\boldsymbol{x}| \, , \; |\boldsymbol{y}^* C_U[h]| \, |diag(H)| = |\boldsymbol{y}^* C_U| \, . \end{array} \tag{6.1}$$

*In addition, in the $\{1,1\}$-case, it is easy to see that:*

$$\begin{aligned} |\boldsymbol{y}^* C_L[q]| \, |A| \, |C_L[p]\boldsymbol{x}| &= |\boldsymbol{y}^* C_L[q]| \begin{bmatrix} 0 & & & & \\ & |a_2| & & & \\ & & \ddots & & \\ & & & |a_{n-1}| & \\ & & & & 0 \end{bmatrix} |C_L[p]\boldsymbol{x}| \\ &= \sum_{i=2}^{n-1} |\boldsymbol{y}^* \left( C_L[q](:,i) a_i C_L[p](i,:) \right) \boldsymbol{x}| \\ &= \sum_{i=2}^{n-1} \left| \boldsymbol{y}^* \begin{bmatrix} 0 & 0 \\ C(i+1:n, 1:i-1) & 0 \end{bmatrix} \boldsymbol{x} \right|, \end{aligned} \tag{6.2}$$

*and analogously,*

$$|\boldsymbol{y}^* C_U[h]| \, |B| \, |C_U[g]\boldsymbol{x}| = \sum_{j=2}^{n-1} \left| \boldsymbol{y}^* \begin{bmatrix} 0 & C(1:j-1, j+1:n) \\ 0 & 0 \end{bmatrix} \boldsymbol{x} \right| . \tag{6.3}$$

*From (6.1), (6.2), and (6.3), we conclude that the respective condition numbers in Theorems 5.2 and 6.2 are the same in the $\{1,1\}$-case.*

On the other hand note that, from the expression obtained in Theorem 6.2 for $\{n_L, n_U\}$-quasiseparable matrices, $\mathrm{cond}(\lambda; \Omega_{QS})$ depends, for values of $n_L$ or $n_u$

greater than 1, on the parameters in the quasiseparable representation $\Omega_{QS}$, and not only on the matrix entries, as it does in the $\{1,1\}$-case.

Another important difference between the $\{1,1\}$-case and higher order cases is that, in general,

$$\operatorname{cond}(\lambda;\Omega_{QS}) \not\leq n\operatorname{cond}(\lambda;C) = n\,\frac{|\boldsymbol{y}^*|\,|C|\,|\boldsymbol{x}|}{|\lambda|\,|\boldsymbol{y}^*\boldsymbol{x}|}, \tag{6.4}$$

where $\operatorname{cond}(\lambda;C)$ is the unstructured eigenvalue condition number in Theorem 3.18, and potentially $\operatorname{cond}(\lambda;\Omega_{QS}) \gg \operatorname{cond}(\lambda;C)$ may happen for $n_L > 1$ or $n_U > 1$. This means that the unstructured condition number can be much smaller than the structured one in those cases, as we have observed in our numerical tests described in Section 6.3.

In order to find a bound for $\operatorname{cond}(\lambda;\Omega_{QS})$ for an $\{n_L, n_U\}$-quasiseparable matrix $C$ with a quasiseparable representation given by $\Omega_{QS}$ as in Theorem 2.23, let us consider the set of all the absolute values of all the parameters in $\Omega_{QS}$ by using the following notation:

$$|\Omega_{QS}| = (\{|p_i|\}_{i=2}^{n}, \{|a_i|\}_{i=2}^{n-1}, \{|q_j|\}_{j=1}^{n-1}, \{|d_i|\}_{i=1}^{n}, \{|g_i|\}_{i=1}^{n-1}, \{|b_i|\}_{i=2}^{n-1}, \{|h_j|\}_{j=2}^{n}),$$

where all $|\cdot|$ must be understood componentwise, and the respective $\{n_L, n_U\}$-quasiseparable matrix $C(|\Omega_{QS}|)$ generated by such parameters.

**Proposition 6.4.** *Let $C$ be an $\{n_L, n_U\}$-quasiseparable matrix with a quasiseparable representation $\Omega_{QS}$. Then,*

$$\operatorname{cond}(\lambda;\Omega_{QS}) \leq n\frac{|\boldsymbol{y}^*|\,C(|\Omega_{QS}|)\,|\boldsymbol{x}|}{|\lambda|\,|\boldsymbol{y}^*\boldsymbol{x}|}.$$

*Proof.* We will proceed by giving a bound for each term in the expression for $\operatorname{cond}(\lambda;\Omega_{QS})$ presented in Theorem 6.2 as follows:

$$\begin{aligned}
\left|\boldsymbol{y}^* \otimes \boldsymbol{e}_{n_L}^T\right| |\operatorname{diag}(P)|\,|C_L[p]\boldsymbol{x}| &= \sum_{i=2}^{n} |y_i|\,|p_i| \left|\begin{bmatrix} a_{i1}^{\times}q_1 & \cdots & a_{i\,i-1}^{\times}q_{i-1} & 0 \end{bmatrix}\boldsymbol{x}\right| \\
&\leq \sum_{i=2}^{n} |y_i|\,C_L(|\Omega_{QS}|)(i,:)\,|\boldsymbol{x}| = |\boldsymbol{y}^*|\,C_L(|\Omega_{QS}|)\,|\boldsymbol{x}|\,, \\
|\boldsymbol{y}^* C_L[q]|\,|\operatorname{diag}(Q)|\,|\boldsymbol{x} \otimes \boldsymbol{e}_{n_L}| &= \sum_{j=1}^{n-1} |\boldsymbol{y}^* \begin{bmatrix} 0 \\ p_{j+1}a_{j+1,j}^{\times} \\ \cdots \\ p_n a_{n,j}^{\times} \end{bmatrix}|\,|q_j|\,|x_j| \\
&\leq \sum_{j=1}^{n-1} |\boldsymbol{y}^*|\,C_L(|\Omega_{QS}|)(:,j)\,|x_j| = |\boldsymbol{y}^*|\,C_L(|\Omega_{QS}|)\,|\boldsymbol{x}|\,,
\end{aligned}$$

$$
\begin{aligned}
\left|\boldsymbol{y}^* \otimes \boldsymbol{e}_{n_U}^T\right| |\mathrm{diag}(G)|\, |C_U[g]\boldsymbol{x}| &= \sum_{i=1}^{n-1} |y_i|\, |g_i| \left|\begin{bmatrix} 0 & b^{\times}_{i\,i+1}h_{i+1} & \cdots & b^{\times}_{i\,n}h_n \end{bmatrix} \boldsymbol{x}\right| \\
&\leq \sum_{i=1}^{n-1} |y_i|\, C_U(|\Omega_{QS}|)(i,:)\, |\boldsymbol{x}| = |\boldsymbol{y}^*|\, C_U(|\Omega_{QS}|)\, |\boldsymbol{x}|\,, \\
|\boldsymbol{y}^* C_U[h]|\, |\mathrm{diag}(H)|\, |\boldsymbol{x} \otimes \boldsymbol{e}_{n_U}| &= \sum_{j=2}^{n} |\boldsymbol{y}^* \begin{bmatrix} g_1 b^{\times}_{i,j} \\ \cdots \\ g_{j-1} b^{\times}_{j-1,j} \\ 0 \end{bmatrix} |\, |h_j|\, |x_j| \\
&\leq \sum_{j=2}^{n} |\boldsymbol{y}^*|\, C_U(|\Omega_{QS}|)(:,j)\, |x_j| = |\boldsymbol{y}^*|\, C_U(|\Omega_{QS}|)\, |\boldsymbol{x}|\,, \\
|\boldsymbol{y}^* C_L[q]|\, |A|\, |C_L[p]\boldsymbol{x}| &= \sum_{i=2}^{n-1} \sum_{l,m=1}^{n_L} \left|\boldsymbol{y}^* \begin{bmatrix} 0 & 0 \\ C[(a_i)_{l\,m}](i+1:n, 1:i-1) & 0 \end{bmatrix} \boldsymbol{x}\right| \\
&\leq \sum_{i=2}^{n-1} |\boldsymbol{y}^*| \begin{bmatrix} 0 & 0 \\ C(|\Omega_{QS}|)(i+1:n, 1:i-1) & 0 \end{bmatrix} |\boldsymbol{x}| \\
&\leq (n-2)|\boldsymbol{y}^*| C_L(|\Omega_{QS}|)|\boldsymbol{x}|, \\
|\boldsymbol{y}^* C_U[h]|\, |B|\, |C_U[g]\boldsymbol{x}| &= \sum_{i=2}^{n-1} \sum_{l,m=1}^{n_U} \left|\boldsymbol{y}^* \begin{bmatrix} 0 & C[(b_i)_{l\,m}](1:i-1, i+1:n) \\ 0 & 0 \end{bmatrix} \boldsymbol{x}\right| \\
&\leq \sum_{i=2}^{n-1} |\boldsymbol{y}^*| \begin{bmatrix} 0 & C(|\Omega_{QS}|)(1:i-1, i+1:n) \\ 0 & 0 \end{bmatrix} |\boldsymbol{x}| \\
&\leq (n-2)|\boldsymbol{y}^*| C_U(|\Omega_{QS}|)|\boldsymbol{x}|,
\end{aligned}
$$

where, in the last expressions above, $C[(a_i)_{l\,m}]$ and $C[(b_i)_{l\,m}]$ are defined as in the proof of Theorem 6.2. Consequently, we have that

$$
\begin{aligned}
\mathrm{cond}(\lambda; \Omega_{QS}) &\leq \frac{1}{|\lambda(\boldsymbol{y}^*\boldsymbol{x})|} \{|\boldsymbol{y}^*|\, C_D(|\Omega_{QS}|)|\boldsymbol{x}| + n\, |\boldsymbol{y}^*|\, C_L(|\Omega_{QS}|)|\boldsymbol{x}| \\
&\qquad\qquad + n\, |\boldsymbol{y}^*|\, C_U(|\Omega_{QS}|)|\boldsymbol{x}|\} \\
&\leq n \frac{|\boldsymbol{y}^*|\, C(|\Omega_{QS}|)\, |\boldsymbol{x}|}{|\lambda|\, |\boldsymbol{y}^*\boldsymbol{x}|}.
\end{aligned}
$$

□

In order to obtain a clear interpretation of Proposition 6.4, we will use the following lemma, which is an extension to rectangular matrices of a result presented in the book by Higham [41, Lemma 3.8].

**Lemma 6.5.** *Let $X_j \in \mathbb{R}^{n_{j-1}\times n_j}$ and $X_j + \Delta X_j \in \mathbb{R}^{n_{j-1}\times n_j}$ be such that $|\Delta X_j| \leq \delta_j |X_j|$ for $j \in \{0, 1, \ldots, m\}$. Then,*

$$\left| \prod_{j=0}^{m} (X_j + \Delta X_j) - \prod_{j=0}^{m} X_j \right| \leq \left( \prod_{j=0}^{m} (1+\delta_j) - 1 \right) \prod_{j=0}^{m} |X_j|.$$

*Proof.* We will proceed by induction. Note first that this result is true for $m = 0$ since:

$$|X_0 + \Delta X_0 - X_0| = |\Delta X_0| \leq \delta_0 |X_0|.$$

Suppose this is a true result for $m-1$ and let us prove it for $m$. Note that

$$\begin{aligned}
\left| \prod_{j=0}^{m} (X_j + \Delta X_j) - \prod_{j=0}^{m} X_j \right| &= \left| \left( \prod_{j=0}^{m-1} (X_j + \Delta X_j) \right) (X_m + \Delta X_m) - \left( \prod_{j=0}^{m-1} X_j \right) X_m \right| \\
&= \left| \left( \prod_{j=0}^{m-1} (X_j + \Delta X_j) - \prod_{j=0}^{m-1} X_j \right) X_m \right. \\
&\quad \left. + \left( \prod_{j=0}^{m-1} (X_j + \Delta X_j) \right) \Delta X_m \right|.
\end{aligned}$$

Then, using the induction hypothesis, we have

$$\begin{aligned}
\left| \prod_{j=0}^{m} (X_j + \Delta X_j) - \prod_{j=0}^{m} X_j \right| &\leq \left( \prod_{j=0}^{m-1} (1+\delta_j) - 1 \right) \prod_{j=0}^{m} |X_j| + \prod_{j=0}^{m-1} (1+\delta_j)\, \delta_m \prod_{j=0}^{m} |X_j| \\
&\leq \left( \prod_{j=0}^{m-1} (1+\delta_j) - 1 + \left( \prod_{j=0}^{m-1} (1+\delta_j) \right) \delta_m \right) \prod_{j=0}^{m} |X_j| \\
&= \left( \prod_{j=0}^{m} (1+\delta_j) - 1 \right) \prod_{j=0}^{m} |X_j|,
\end{aligned}$$

which completes the proof. □

Using Lemma 6.5, we can bound the perturbations of the $\{n_L, n_U\}$-quasiseparable matrix $C$ under tiny relative perturbations on the parameters in the quasiseparable representation $\Omega_{QS}$ described in Theorem 2.23. We denote such perturbations on the parameters as:

$$\begin{aligned}
\Omega_{QS} + \delta\Omega_{QS} = (&\{p_i + \delta p_i\}_{i=2}^{n}, \{a_i + \delta a_i\}_{i=2}^{n-1}, \{q_j + \delta q_j\}_{j=1}^{n-1}, \{d_i + \delta d_i\}_{i=1}^{n}, \\
&\{g_i + \delta g_i\}_{i=1}^{n-1}, \{b_i + \delta b_i\}_{i=2}^{n-1}, \{h_j + \delta h_j\}_{j=2}^{n}),
\end{aligned}$$

and we will consider $|\delta\Omega_{QS}| \leq \eta\, |\Omega_{QS}|$. i.e.,

$$\begin{aligned}
&|\delta p_i| \leq \eta |p_i|,\ |\delta a_i| \leq \eta |a_i|,\ |\delta q_j| \leq \eta |q_j|,\ |\delta d_i| \leq \eta |d_i|, \\
&|\delta g_i| \leq \eta |g_i|,\ |\delta b_i| \leq \eta |b_i|,\ |\delta h_j| \leq \eta |h_j|.
\end{aligned}$$

**Lemma 6.6.** *With the notation above, we have*

$$|C(\Omega_{QS} + \delta\Omega_{QS}) - C(\Omega_{QS})| \leq [(1+\eta)^n - 1]\, C(|\Omega_{QS}|).$$

*Proof.* Since each entry of $C(\Omega_{QS})$ is a product of at most $n$ matrices, as a consequence of Lemma 6.5, we have that

$$\left| [C(\Omega_{QS} + \delta\Omega_{QS})]_{i\,j} - [C(\Omega_{QS})]_{i\,j} \right| \leq [(1+\eta)^n - 1]\, [C(|\Omega_{QS}|)]_{i\,j}\,.$$

□

Note that the bound in Lemma 6.6 is not always attained for all entries, but it is the best that we can say for all the entries of $C$. The key point to this matter is that the parameters in $\Omega_{QS}$ may not determine well the matrix under tiny perturbations if $C(|\Omega_{QS}|)_{ij} \gg |C(\Omega_{QS})_{ij}|$ for some entry $C(\Omega_{QS})_{ij}$ not far from $\max_{kl} |C(\Omega_{QS})_{kl}|$.

These comments allow us to give a clear interpretation of Proposition 6.4, since Lemma 6.6 suggests the use of the *unstructured condition number* introduced by Higham and Higham in [42], with respect to $C(|\Omega_{QS}|)$:

$$\begin{aligned}\mathrm{cond}_{|\Omega_{QS}|}(\lambda; C) &:= \lim_{\eta\to 0} \sup \left\{ \frac{|\delta\lambda|}{\eta|\lambda|} \,:\, (\lambda+\delta\lambda) \text{ is eigenvalue of } (C+\delta C), \right. \\ &\qquad\qquad \left. |\delta C| \leq \eta |C\left(|\Omega_{QS}|\right)| \right\} \\ &= \frac{|\boldsymbol{y}^*|\, C\left(|\Omega_{QS}|\right) |\boldsymbol{x}|}{|\lambda|\, |\boldsymbol{y}^*\boldsymbol{x}|},\end{aligned}$$

which is, essentially, the bound given for $\mathrm{cond}(\lambda; \Omega_{QS})$ in Proposition 6.4, up to a factor $n$.

**Remark 6.7.** *The condition number $\mathrm{cond}_{|\Omega_{QS}|}(\lambda; C)$, defined above, is also a particular case of the general condition number in Definition 3.24. This is observed by considering $\Omega = C$ and $E = |\Omega_{QS}|$ in Definition 3.24, i.e.: $\mathrm{cond}_{|\Omega_{QS}|}(\lambda; C) = \mathrm{cond}_{|\Omega_{QS}|}(\lambda, C; C)$.*

On the other hand, it is easy to check that the relative condition number $\mathrm{cond}(\lambda; \Omega_{QS})$ for $\{n_L, n_U\}$-quasiseparable matrices is also invariant under diagonal similarities.

**Proposition 6.8.** *Let $K = \mathrm{diag}\,(k_1, k_2, \cdots, k_n)$ be an invertible diagonal matrix and $C \in \mathbb{R}^{n\times n}$ be an $\{n_L, n_U\}$-quasiseparable matrix. Then, the following assertions hold.*

*(a) The matrix $KCK^{-1} \in \mathbb{R}^{n\times n}$ is also an $\{n_L, n_U\}$-quasiseparable matrix.*

*(b) If the set of parameters*

$$\Omega_{QS} = (\{p_i\}_{i=2}^{n}, \{a_i\}_{i=2}^{n-1}, \{q_i\}_{i=1}^{n-1}, \{d_i\}_{i=1}^{n}, \{g_i\}_{i=1}^{n-1}, \{b_i\}_{i=2}^{n-1}, \{h_i\}_{i=2}^{n}),$$

*defines a quasiseparable representation of $C$, then the set of parameters*

$$\begin{aligned}\Omega'_{QS} &= (\{p'_i\}_{i=2}^{n}, \{a'_i\}_{i=2}^{n-1}, \{q'_i\}_{i=1}^{n-1}, \{d'_i\}_{i=1}^{n}, \{g'_i\}_{i=1}^{n-1}, \{b'_i\}_{i=2}^{n-1}, \{h'_i\}_{i=2}^{n}) \\ &= (\{k_i p_i\}_{i=2}^{n}, \{a_i\}_{i=2}^{n-1}, \{q_i/k_i\}_{i=1}^{n-1}, \{d_i\}_{i=1}^{n}, \{k_i g_i\}_{i=1}^{n-1}, \{b_i\}_{i=2}^{n-1}, \{h_i/k_i\}_{i=2}^{n}),\end{aligned}$$

*defines a quasiseparable representation of $KCK^{-1}$.*

*(c) For the quasiseparable representations $\Omega_{QS}$ and $\Omega'_{QS}$ of the matrices $C$ and $KCK^{-1}$ respectively defined above, the relative eigenvalue condition number is invariant, i.e.,*

$$\mathrm{cond}(\lambda, C; \Omega_{QS}) = \mathrm{cond}(\lambda, KCK^{-1}; \Omega'_{QS}).$$

*Proof.* The assertions in $(a)$ and $(b)$ are proved just as in Lemma 5.8 for $\{1,1\}$-quasiseparable matrices, while the invariance under diagonal similarities for $\mathrm{cond}(\lambda; \Omega_{QS})$ stated in $(c)$, is a direct consequence of $(a)$ and $(b)$. Let $C' = KCK^{-1}$ and consider $\Omega'_{QS}$ as in $(b)$. Then, using $(d)$ from Lemma 5.7, we have that if $\boldsymbol{y}$ and $\boldsymbol{x}$ are left and right eigenvectors of $C$, respectively, then $\boldsymbol{y}_{C'} = (\boldsymbol{y}^* K^{-1})^*$ and $\boldsymbol{x}_{C'} = K\boldsymbol{x}$ are left and right eigenvectors of $C'$, respectively. Furthermore, from the proof of Theorem 5.9, we have $|\boldsymbol{y}^*\boldsymbol{x}| = |\boldsymbol{y}_{C'}^*\boldsymbol{x}_{C'}|$, and since $\{a'_i\}_{i=2}^{n-1} = \{a_i\}_{i=2}^{n-1}$, $\{d'_i\}_{i=1}^{n} = \{d_i\}_{i=1}^{n}$, and $\{b'_i\}_{i=2}^{n-1} = \{b_i\}_{i=2}^{n-1}$, we only need to compare the respective contributions of the other parameters in $\Omega_{QS}$ and $\Omega'_{QS}$ to $\mathrm{cond}(\lambda, C; \Omega_{QS})$ and $\mathrm{cond}(\lambda, C'; \Omega'_{QS})$. We will only develop the comparison for the parameters representing the lower triangular parts of $C$ and $C'$ because the comparison for the parameters representing the respective upper triangular parts of these matrices can be done in an analogous way. If, for the matrix $C'$ and the representation $\Omega'_{QS}$, we define $C'_L[p]$, $P'$, $C'_L[q]$, and $Q'$ in an analogous way to $C_L[p]$, $P$, $C_L[q]$, and $Q$, for the matrix $C$ and the representation $\Omega_{QS}$, as in Definition 6.1, then, using the expression in Theorem 2.23 for representing $\{n_L, n_U\}$-quasiseparable matrices, we have

$$\begin{aligned}|\boldsymbol{y}^* \otimes \boldsymbol{e}_{n_L}^T||diag(P)||C_L[p]\boldsymbol{x}| &= |(\boldsymbol{y}_{C'}^* K) \otimes \boldsymbol{e}_{n_L}^T||diag(P)||C_L[p](K^{-1}\boldsymbol{x}_{C'})| \\ &= |(\boldsymbol{y}_{C'}^* \otimes \boldsymbol{e}_{n_L}^T)(K \otimes I_{n_L})||diag(P)||(C_L[p]K^{-1})\boldsymbol{x}_{C'}| \\ &= |\boldsymbol{y}_{C'}^* \otimes \boldsymbol{e}_{n_L}^T||(K \otimes I_{n_L})diag(P)||(C'_L[p])\boldsymbol{x}_{C'}| \\ &= |\boldsymbol{y}_{C'}^* \otimes \boldsymbol{e}_{n_L}^T||diag(P')||(C'_L[p])\boldsymbol{x}_{C'}|,\end{aligned}$$

and

$$|\boldsymbol{y}^* C_L[q]||diag(Q)||\boldsymbol{x} \otimes \boldsymbol{e}_{n_L}| = |(\boldsymbol{y}_{C'}^* K)C_L[q]||diag(Q)||(K^{-1}\boldsymbol{x}_{C'}) \otimes \boldsymbol{e}_{n_L}|$$

$$
\begin{aligned}
&= |\boldsymbol{y}^*_{C'}(KC_L[q])||diag(Q)||(K^{-1} \otimes I_{n_L})(\boldsymbol{x}_{C'} \otimes \boldsymbol{e}_{n_L})| \\
&= |\boldsymbol{y}^*_{C'}C'_L[q]||diag(Q)(K^{-1} \otimes I_{n_L})||(\boldsymbol{x}_{C'} \otimes \boldsymbol{e}_{n_L})| \\
&= |\boldsymbol{y}^*_{C'}C'_L[q]||diag(Q')||(\boldsymbol{x}_{C'} \otimes \boldsymbol{e}_{n_L})|,
\end{aligned}
$$

where we have used the mixed-product property of the Kronecker product, that is, $(A \otimes B)(C \otimes D) = (AC) \otimes (BD)$, for any matrices $A$, $B$, $C$ and $D$ such that one can form the matrix products $AC$ and $BD$ [44].

The proof follows by considering the expression in Theorem 6.2 for $\text{cond}(\lambda, C; \Omega_{QS})$ and $\text{cond}(\lambda, KCK^{-1}; \Omega'_{QS})$. □

## 6.2 Fast computation of the eigenvalue condition number in the quasiseparable representation

In this section, we prove that the eigenvalue condition number $\text{cond}(\lambda; \Omega_{QS})$, presented in Theorem 6.2 for general $\{n_L, n_U\}$-quasiseparable matrices, can be computed fast as in the $\{1, 1\}$-case. Furthermore, in the proof of Proposition 6.9, we provide an algorithm for its computation.

**Proposition 6.9.** *Let $C \in \mathbb{R}^{n \times n}$ be an $\{n_L, n_U\}$-quasiseparable matrix with a simple eigenvalue $\lambda \neq 0$ with left eigenvector $\boldsymbol{y}$ and right eigenvector $\boldsymbol{x}$. Assume that $\lambda, \boldsymbol{x}, \boldsymbol{y}$, and the quasiseparable representation $\Omega_{QS}$ of $C$ are all known. Then, $\text{cond}(\lambda; \Omega_{QS})$ can be computed in $\mathcal{O}((n_L^2 + n_U^2)n)$ flops.*

*Proof.* We already know from Proposition 5.10 that the factor $|\lambda(\boldsymbol{y}^*\boldsymbol{x})|^{-1}$ and the term $|\boldsymbol{y}^*|\,|C_D|\,|\boldsymbol{x}|$ in the expression in Theorem 6.2 for $\text{cond}(\lambda; \Omega_{QS})$ can both be computed in $\mathcal{O}(n)$ flops. Therefore, we will only focus on the main cost of computing the other terms in Theorem 6.2 as follows.

(a) In the case of the term $|\boldsymbol{y}^* \otimes \boldsymbol{e}_{n_L}^T||\text{diag}(P)||C_L[p]\boldsymbol{x}|$, the main cost comes from computing the product $C_L[p]\boldsymbol{x}$. For such a purpose, we can use a similar algorithm to the one used in the proof of Proposition 5.10. Consider the vector $z = C_L[p]\boldsymbol{x}$, where $z$ is partitioned in $n$ block rows of size $n_L$ each. Recall that in the $i$th position of $z$ there is a vector $z_i$ of size $n_L \times 1$. The fast method for computing $z$ is described in the following routine.

**Routine 2** (Computes $z = C_L[p]\boldsymbol{x}$.)

$z_1 = 0$

$z_2 = q_1 x_1$

`for` $i = 3 : n$

$z_i = a_{i-1} z_{i-1} + q_{i-1} x_{i-1}$

`endfor`

For checking the validity of this algorithm for finding $\boldsymbol{z}$, note first that $\boldsymbol{z}_1$ and $\boldsymbol{z}_2$ are found correctly. Then, by using induction, suppose that Routine 2 has computed correctly $z_1, z_2, \ldots, z_i$ and let us check that it computes correctly $z_{i+1}$. Observe that at step $i$ we have

$$z_i = a_{i-1}a_{i-2}\cdots a_2 q_1 x_1 + a_{i-1}a_{i-2}\cdots a_3 q_2 x_2 + \cdots + a_{i-1}q_{i-2}x_{i-2} + q_{i-1}x_{i-1}.$$

Therefore, at step $i+1$, we have

$$\begin{aligned} a_i z_i + q_i x_i &= a_i a_{i-1}a_{i-2}\cdots a_2 q_1 x_1 + a_i a_{i-1}a_{i-2}\cdots a_3 q_2 x_2 + \cdots + a_i a_{i-1}q_{i-2}x_{i-2} \\ &\quad + a_i q_{i-1}x_{i-1} + q_i x_i \\ &= z_{i+1}, \end{aligned}$$

which is also correct.
Since $a_i$ is a real matrix of size $n_L \times n_L$ and $q_i$ is a real column vector of size $n_L$, using Routine 2 for computing the product $C_L[p]\boldsymbol{x}$, it is easy to see that the cost of computing the term in $(a)$ is $\mathcal{O}(n_L^2 n)$ flops.

(b) For the term $|\boldsymbol{y}^* C_L[q]||\mathrm{diag}(Q)||\boldsymbol{x} \otimes \boldsymbol{e}_{n_L}|$, we can use a similar procedure to that in (a), and compute it also with a cost of $\mathcal{O}(n_L^2 n)$ flops.

(c) The terms $|\boldsymbol{y}^* \otimes \boldsymbol{e}_{n_U}^T||\mathrm{diag}(G)||C_U[g]\boldsymbol{x}|$ and $|\boldsymbol{y}^* C_U[h]||\mathrm{diag}(H)||\boldsymbol{x} \otimes \boldsymbol{e}_{n_U}|$ can both be computed in an analogous way to the terms in $(a)$ and $(b)$, respectively, with a cost of $\mathcal{O}(n_U^2 n)$ flops.

(d) The term $|\boldsymbol{y}^* C_L[q]|\,|A|\,|C_L[p]\boldsymbol{x}| = \sum_{i=2}^{n-1} |\boldsymbol{y}^* C_L[q](:,i)|\,|a_i|\,|C_L[p](i,:)\boldsymbol{x}|$ can obviously be computed in $\mathcal{O}(n_L^2 n)$ flops, since the products $\boldsymbol{y}^* C_L[q]$ and $C_L[p]\boldsymbol{x}$ have been already respectively computed in (b) and (a).

(e) The term $|\boldsymbol{y}^* C_U[h]|\,|B|\,|C_U[g]\boldsymbol{x}| = \sum_{i=2}^{n-1} |\boldsymbol{y}^* C_U[h](:,i)|\,|b_i|\,|C_U[g](i,:)\boldsymbol{x}|$ can also be computed in $\mathcal{O}(n_U^2 n)$ flops, since the products $\boldsymbol{y}^* C_U[h]$ and $C_U[g]\boldsymbol{x}$ have been already computed in (c).

Finally, we prove the desired result by summing all the costs obtained above. □

**Remark 6.10.** *Note that for $\{1,1\}$-quasiseparable matrices we can also use the procedure described in the proof of Proposition 6.9 for computing the parameterized eigenvalue condition number. In particular, computing, in the way described above, the terms in $(d)$ and $(e)$ from Proposition 6.9 can simplify the more involved computations described in $(g)$ and $(h)$ from Proposition 5.10 for the $\{1,1\}$-case.*

## 6.3 Numerical experiments

In this section we provide a brief description of some numerical experiments that have been performed in order to compare the structured eigenvalue condition number

$\text{cond}(\lambda; \Omega_{QS})$ presented in Theorem 6.2 for a general quasiseparable representation of an $\{n_L, n_U\}$-quasiseparable matrix $C$, with the unstructured one $\text{cond}(\lambda; C)$ in Definition 3.17 and computed using Theorem 3.18. We have used MATLAB for running several random numerical tests. First, the command `randn` from MATLAB has been used for generating the random parameters in a quasiseparable representation for an $\{n_L, n_U\}$-quasiseparable matrix of size $n \times n$, i.e., the following random vectors and matrices of parameters are generated:

$$\begin{aligned} &\boldsymbol{d} \in \mathbb{R}^{1\times n}, \{\boldsymbol{p_i}\}_{i=2}^{n} \subset \mathbb{R}^{1\times n_L}, \{\boldsymbol{a_i}\}_{i=2}^{n-1} \subset \mathbb{R}^{n_L\times n_L}, \{\boldsymbol{q_j}\}_{j=1}^{n-1} \subset \mathbb{R}^{n_L\times 1}, \\ &\{\boldsymbol{g_i}\}_{i=1}^{n-1} \subset \mathbb{R}^{1\times n_U}, \{\boldsymbol{b_i}\}_{i=2}^{n-1} \subset \mathbb{R}^{n_U\times n_U}, \{\boldsymbol{h_j}\}_{j=2}^{n} \subset \mathbb{R}^{n_U\times 1}. \end{aligned} \tag{6.5}$$

Then, we build the quasiseparable matrix $C$ of size $n \times n$ generated by these parameters and the quasiseparable matrix $C(|\Omega_{QS}|)$ of size $n \times n$ generated by the absolute values of these parameters. We obtain the eigenvalues and eigenvectors of $C$ using the standard command `eig` from MATLAB. Finally, we compute the structured eigenvalue condition number $\text{cond}(\lambda; \Omega_{QS})$, using the fast Algorithm described in Section 6.2, and the unstructured condition numbers $\text{cond}(\lambda; C)$ and $\text{cond}_{|\Omega_{QS}|}(\lambda; C)$, using direct matrix-vector multiplication with a resulting cost of order $2n^2 + \mathcal{O}(n)$ operations each (note that $\text{cond}(\lambda; C)$ and $\text{cond}_{|\Omega_{QS}|}(\lambda; C)$ could also be computed in $\mathcal{O}(n)$ operations each since $C$ and $C(|\Omega_{QS}|)$ are also quasiseparable matrices).

In order to verify the inequality (6.4) in Section 6.1, we compare the condition numbers $\text{cond}(\lambda; \Omega_{QS})$ and $\text{cond}(\lambda; C)$ obtaining similar, often moderate, values, i.e., $\text{cond}(\lambda; \Omega_{QS}) \approx \text{cond}(\lambda; C)$, for strictly random generated parameteres as in (6.5). Therefore, in order to maximize de quotient $(\text{cond}(\lambda; \Omega_{QS}))/(\text{cond}(\lambda; C))$, we used the multidirectional search method for direct search optimization `mdsmax` from the Matrix Computation Toolbox (see [41, Appendix D]), using as input a linear vector **lpar** constructed by concatenating all vectors and matrices in $\Omega_{QS}$ reshaped as linear vectors by using the `reshape` function from MATLAB. In this way, we obtained particular sets of quasiseparable parameters that generated matrices with simple eigenvalues for which $\text{cond}(\lambda; C) \ll \text{cond}(\lambda; \Omega_{QS})$. In those cases we also computed the condition number $\text{cond}_{|\Omega_{QS}|}(\lambda; C)$. In fact, for $(n, n_L, n_U) = (10, 2, 2)$ and $(n, n_L, n_U) = (15, 3, 3)$, we obtained the following results:

| $(n, n_L, n_U)$ | $\lambda$ | $\text{cond}(\lambda; \Omega_{QS})$ | $\text{cond}(\lambda; C)$ | $\text{cond}_{\lvert\Omega_{QS}\rvert}(\lambda; C)$ |
|---|---|---|---|---|
| $(10, 2, 2)$ | $2.9781 \cdot 10^2 - 6.7413 \cdot 10^2 i$ | $1.6910 \cdot 10^5$ | 1.7827 | $1.9687 \cdot 10^5$ |
| $(15, 3, 3)$ | $1.3244 \cdot 10^5 - 3.4536 \cdot 10^5 i$ | $7.9502 \cdot 10^5$ | 1.3455 | $9.0614 \cdot 10^6$ |

.

These examples show not only that the structured eigenvalue condition number $\text{cond}(\lambda; \Omega_{QS})$ may be much larger than the unstructured one $\text{cond}(\lambda; C)$, but also that there exist $\{n_L, n_U\}$-quasiseparable matrices having eigenvalues that are very well conditioned with respect to relative entrywise perturbations on the matrix but that are ill conditioned with respect to relative componentwise perturbations on the parameters of certain quasiseparable representations of the matrix. In addition,

note that this large differences occurred when $\mathrm{cond}(\lambda; C) \ll \mathrm{cond}_{|\Omega_{QS}|}(\lambda; C)$. This fact, reinforces our interpretation of Proposition 6.4, given in the paragraphs below Lemma 6.6.

# Chapter 7

# Conclusions, publications, and open problems

In this chapter we summarize the main results and original contributions of this dissertation. We discuss some future work motivated by this thesis, list the published papers including most of the original results presented in this thesis, and the conferences where they have been presented.

## 7.1 Conclusions and original contributions

A summary of the main original results in the chapters of this thesis is provided below.

**Chapter 3:** A general expression for the condition number of the solution of a linear system of equations whose coefficient matrix is a differentiable function of a vector of parameters with respect to relative componentwise perturbations of such parameters has been presented. This expression involves the partial derivatives of the matrix with respect to the parameters.

In addition, we have also provided a general formula for the relative condition number of a simple eigenvalue of any matrix that can be represented by a vector of parameters, in a differentiable way, with respect to relative componentwise perturbations of these parameters. The results in this chapter can be generalized to other conditioning problems and matrices depending on parameters.

**Chapter 4:** The general expression introduced in Section 3.1.2 from Chapter 3, for the condition number of the solution of a linear system of equations whose coefficient matrix is a differentiable function of a vector of parameters with respect to relative componentwise perturbations of such parameters, has been used to deduce formulas for the componentwise condition numbers of the solutions of linear systems whose coefficient matrices are $\{1, 1\}$-quasiseparable

of size $n \times n$ with respect to perturbations of the parameters in any quasiseparable representation and in the tangent-Givens-vector representation of the coefficient matrices. We have compared theoretically these two structured condition numbers and we have proved that they differ at most by a factor $3n$ and, therefore, that they are numerically equivalent, though the one with respect to the tangent-Givens-vector representation is always the smallest. Moreover, it has been shown that these structured condition numbers can be estimated in $\mathcal{O}(n)$ operations via an effective condition number. We have also proved rigorously that these structured condition numbers are always smaller, up to a factor $n$, than the componentwise unstructured condition number. In addition, the performed numerical experiments illustrate that the structured condition numbers can be much smaller than the unstructured one in practice. This means that the structure of $\{1,1\}$-quasiseparable matrices may play a key role in the accuracy of computed solutions of linear systems of equations, since these solutions can be much less sensitive to relative perturbations of the parameters representing the matrices than to relative perturbations of the matrix entries. The techniques used in this chapter can be generalized to obtain structured condition numbers for the solution of linear systems involving other classes of low-rank structured matrices and they can be extended to study the structured conditioning of other problems involving low-rank structured matrices like, for instance, least squares problems.

**Chapter 5:** Based on the formula introduced in Section 3.2.2 from Chapter 3, we have obtained expressions for the eigenvalue condition numbers of $\{1,1\}$-quasiseparable matrices with respect to relative componentwise perturbations of the parameters in the quasiseparable and the Givens-vector representations, and we have developed fast algorithms with cost $O(n)$ operations for computing these condition numbers. As far as we know, these results are the first ones available in the literature dealing with structured perturbations of low-rank structured matrices (the results for the structured conditioning of eigenvalues in this chapter where obtained and published before those in Chapter 4 for linear systems, but we decided to describe them in the reverse order in this dissertation for keeping consistency with the order in most of the books in Linear algebra, that is, the results for linear systems are explained before those for eigenvalues). Numerical tests comparing the new structured eigenvalue condition numbers with the unstructured componentwise condition number have been performed and have revealed that the eigenvalues of quasiseparable matrices may be very well-conditioned under relative perturbations of the parameters, but very ill-conditioned under general unstructured relative perturbations of the matrix entries. In contrast, it has been proved theoretically that the opposite cannot happen. Therefore, for $\{1,1\}$-quasiseparable matrices, we have established that the structure should play a key role in the accuracy of eigenvalue computations since the sensitivity of their simple eigenvalues is potentially much

smaller to perturbations of the parameters in the covered representations than to perturbations of the matrix entries. In addition, we have proved that all considered representations of $\{1,1\}$-quasiseparable matrices are equivalent with respect to the eigenvalue sensitivity and that the Givens-vector representation is the one leading to the smallest eigenvalue condition numbers.

**Chapter 6:** Using the formula for the eigenvalue condition number for parameterized matrices given in Section 3.2.2 from Chapter 3, we have obtained an explicit expression for the condition number for a simple eigenvalue of an $\{n_L, n_U\}$-quasiseparable matrix with respect to relative perturbations on the parameters in the general quasiseparable representation that generates the matrix. We have also proved that this condition number can be computed fast via an algorithm of cost $\mathcal{O}((n_L^2 + n_U^2)n)$ flops. On the other hand, for the general $\{n_L, n_U\}$-case, significant differences have been observed with respect to the $\{1,1\}$-case studied in Chapter 5, for the quasiseparable representation. In fact, we noted that if $n_L > 1$ or $n_U > 1$ then the structured eigenvalue condition number depends on the parameters in the quasiseparable representation of the matrix and, what it is more important, this structured condition number can be significantly larger than the unstructured one, as we have observed in our numerical experiments described in Section 6.3. Consequently, we have provided an interpretation of such differences, and suggested the use of an unstructured condition number given in terms of the entries of the quasiseparable matrix generated by the absolute values of the parameters in the general quasiseparable representation of the original matrix. To summarize, we have proved that describing the structure of a quasiseparable matrix with a high order of quasiseparability by using the general quasiseparable representation may lead to much larger eigenvalue condition numbers than the unstructured ones. This suggests the convenience of studying a different way of describing high order quasiseparable structures, i.e., to study other representations for such matrices, in order to exploit the structure and obtain smaller condition numbers. This remains as a future research project.

## 7.2 Publications

The results in Section 3.1.2 and Chapter 4 are contained in:

Dopico, F. M., Pomés, K., *Structured condition numbers for linear systems with parameterized quasiseparable coefficient matrices*, published electronically in Numerical Algorithms, DOI 10.1007/s11075-016-0133-8, 2016.

The results in Section 3.2.2 and Chapter 5 are contained in:

> Dopico, F. M., Pomés, K., *Structured eigenvalue condition numbers for parameterized quasiseparable matrices*, published electronically in Numerische Mathematik, DOI 10.1007/s00211-015-0779-5, 2015.

## 7.3 Contributions to Conferences

The results in this dissertation were described by its author in the following presentations:

*Structured condition numbers for the solution of parameterized quasiseparable linear systems.* Presented as a contributed talk in the Society for Industrial and Applied Mathematics (SIAM) Conference on Applied Linear Algebra, October 26-30, Atlanta, USA. In this talk, most of the original results from Chapters 3 and 4 on structured condition numbers for linear systems with $\{1,1\}$-quasiseparable matrices of coefficients were presented.

*Parameterized condition numbers for the solution of quasiseparable linear systems.* Presented in the Young Researchers Sessions in the Red de Álgebra Lineal, Análisis Matricial y Aplicaciones (ALAMA) Meeting, June 20-22, León, Spain. In this talk, most of the original results from Chapters 3 and 4 on condition numbers for linear systems with $\{1,1\}$-quasiseparable matrices of coefficients were also presented.

*Parameterized condition numbers for quasiseparable matrices.* Presented as a contributed talk in the 20-th Conference of the International Linear Algebra Society (ILAS), July 11-15, KU Leuven, Belgium. In this talk, most of the original results obtained for parameterized condition numbers for $\{1,1\}$-quasiseparable matrices, included in Chapters 3, 4, and 5, were presented together with a comparison on the results for the condition number for eigenvalues in the general $\{n_L, n_U\}$-case, included in Chapter 6 of this thesis. Some partial results on eigenvalue condition numbers for $\{n_L, n_U\}$-quasiseparable matrices in the Givens-weight representation, which are not included in this dissertation, were also commented.

## 7.4 Future Work

In this section, we provide a description of some open problems related to the results included in this dissertation.

One of the main objectives of this work was to establish a framework for comparing the quasiseparable and the Givens-vector representations in terms of the

sensitivity of the solutions of different problems involving quasiseparable matrices. For the $\{1,1\}$-case, this comparison has been completed for simple eigenvalues of quasiseparable matrices and for the solution of linear systems of equations with a quasiseparable matrix of coefficients, by proving that the Givens-vector representation produces smaller condition numbers than the quasiseparable one, but that there are not significant relative differences among them, and that, therefore, they can be considered numerically equivalent. On the other hand, for the general $\{n_L, n_U\}$-case we have only covered the eigenvalue condition number in the general quasiseparable representation. In this case we have proved that this structured condition number can be much larger than the unstructured one, and , probably, this will also happen for the solution of linear systems because this representation might not represent well the entries of the matrix under certain circumstances, as we commented in Section 6.1, this obviosly suggests the use of a different and more stable representation. In [17], a generalization of the Givens-vector representation for high orders of quasiseparability, named the Givens-weight representation, was introduced. That representation can be used, for instance, for computing the Hessenberg reduction, the $QR$-factorization, the solution of linear systems, and the eigenvalues of $\{n_L, n_U\}$-quasiseparable matrices, as described in [18, 19, 20]. Therefore, the first step in the future is to extend the results covered in this work to general quasiseparable matrices in the Givens-weight representation. In particular, we expect to obtain structured condition numbers for linear systems ans eigenvalues in the Givens-weight representation, using the symbolic formalism introduced in the recent and still unpublished paper [57] for describing sequences of transformations instead of the more graphical approaches in some previous works [17, 54, 63]. In a second step, we plan to compare these condition numbers in the Givens-weight representation with the unstructured ones and to prove that they are never much larger than the unstructured ones, but that they can be much smaller.

In other context, there are still many problems in numerical linear algebra involving rank structured matrices (or more in general, parameterized matrices) for which the structured conditioning of the solutions may be studied. In particular, it would be very interesting to extend the results obtained in Chapters 4 and 5 for linear systems and eigenvalues, respectively, to other classical problems in this area and obtain structured condition numbers for singular values, eigenvectors and the solutions of least squares problems for quasiseparable matrices. In order to achieve this, the general approach in Sections 3.1.2 and 3.2.2 for parameterized condition numbers must be extended to those classical problems.

# Bibliography


[1] Asplund, E., *Inverses of matrices $a_{ij}$ which satisfy $a_{ij} = 0$ for $j > i + p$*, Math. Scand., 7:57-60, 1959.

[2] Asplund, S.O., *Finite boundary value problems solved by Green's matrix*, Math. Scand., 7:49-56, 1959.

[3] Aurentz, J.L., Mach, T., Vandebril, R., Watkins, D.S., *Fast and backward stable computation of roots of polynomials*, SIAM J. Matrix Anal. Appl., 36(3):942-973, 2015.

[4] Barlow, J.L., Ipsen, I.C.F., *Scaled Givens rotations for the solution of linear least squares problems on systolic arrays*, SIAM J. Sci. Stat. Comput., 8(5):716-733, 1987.

[5] Bebendorf, M., *Why finite element discretizations can be factored by triangular hierarchical matrices*, SIAM J. Numer. Anal., 45(4):1472-1494, 2007.

[6] Bella, T., Olshevsky, V., Stewart, M., *Nested product decomposition of quasiseparable matrices*, SIAM J. Matrix Anal. Appl., 34(4):1520-1555, 2013.

[7] Berger, W.J., Saibel, E., *On the inversion of continuant matrices*, Frankl. Inst.-Eng Appl. Math., 6:249-253, 1953.

[8] Bindel, D., Demmel, J.W., Kahan, W., Marques, O., *On computing Givens rotations reliably and efficiently*, ACM Trans. Math. Softw., 28(2):206-238, 2002.

[9] Bini, D.A., Gemignani, L., Pan, V.Y., *Fast and stable QR eigenvalue algorithms for generalized companion matrices and secular equations*, Numer. Math., 100(3):373-408, 2005.

[10] Bini, D.A., Gemignani, L., Tisseur, F., *The Ehrlich-Aberth method for the nonsymmetric tridiagonal eigenvalue problem*, SIAM J. Matrix Anal. Appl., 27(1):153-175, 2005.

[11] Bini, D.A., Boito, P., Eidelman, Y., Gemignani, L., Gohberg, I., *A fast implicit QR eigenvalue algorithm for companion matrices*, Linear Algebra Appl., 432:2006-2031, 2010.

[12] Börm, S., Grasedyck, L., *Hybrid cross approximation of integral operators*, Numer. Math., 101:221-249, 2005.

[13] Börm, S., Grasedyck, L., Hackbusch, W., *Introduction to hierarchical matrices with applications*, Eng. Anal. Bound. Elemen., 27:405-422, 2003.

[14] Chandrasekaran, S., Gu, M., Xia, J., Zhu, J., *A fast QR algorithm for companion matrices*, Recent advances in matrix and operator theory, 111-143, Oper. Theory Adv. Appl., 179, Birkhäuser, Basel, 2008.

[15] Chandrasekaran, S., Gu, M., *Fast and stable eigendecomposition of symmetric banded plus semiseparable matrices*, Linear Algebra Appl., 313:107-114, 2000.

[16] Chandrasekaran, S., Mastronardi, M., Van Huffel, S., *Fast and stable algorithms for reducing matrices to tridiagonal and bidiagonal form*, BIT 41, 1:7-12, 2003.

[17] Delvaux, S., Van Barel, M., *A Givens-weight representation for rank structured matrices*, SIAM J. Matrix Anal. Appl., 29(4):1147-1170, 2007.

[18] Delvaux, S., Van Barel, M., *A Hessenberg reduction algorithm for rank structured matrices*, SIAM J. Matrix Anal. Appl., 29(3):895-926, 2007.

[19] Delvaux, S., Van Barel, M., *Eigenvalue computation for unitary rank structured matrices*, J. Comput. Appl. Math., 213:268-287, 2008.

[20] Delvaux, S., Van Barel, M., *A QR-based solver for rank structured matrices*, SIAM J. Matrix Anal. Appl., 30(2):464-490, 2008.

[21] Dopico, F.M., Olshevsky, V., Zhlobich, P., *Stability of QR-based fast system solvers for a subclass of quasiseparable rank one matrices*, Math. Comp., 82:2007-2034, 2013.

[22] Dopico, F.M., Pomés, K., *Structured eigenvalue condition numbers for parameterized quasiseparable matrices*, published electronically in Numer. Math., DOI 10.1007/s00211-015-0779-5, 2015.

[23] Dopico, F.M., Pomés, K., *Structured condition numbers for linear systems with parameterized quasiseparable coefficient matrices*, published electronically in Numer. Algor., DOI 10.1007/s11075-016-0133-8, 2016.

[24] Eidelman, Y., Gohberg, I., *On a new class of structured matrices*, Integral Equ. Oper. Theory, 34(3):293-324, 1999.

[25] Eidelman, Y., Gohberg, I., *Linear complexity inversion algorithms for a class of structured matrices*, Integral Equ. Oper. Theory, 35:28-52, 1999.

[26] Eidelman, Y., Gohberg, I., *On generators of quasiseparable finite block matrices*, Calcolo, 42:187-214, 2005.

[27] Eidelman, Y., Gohberg, I., Haimovici, I., *Separable Type Representations of Matrices and Fast Algorithms. Volume 1. Basics. Completion Problems. Multiplication and Inversion Algorithms*, Operator Theory: Advances and Applications, 234. Birkhäuser/Springer, Basel, 2014.

[28] Eidelman, Y., Gohberg, I., Haimovici, I., *Separable Type Representations of Matrices and Fast Algorithms. Volume 2. Eigenvalue Method*, Operator Theory: Advances and Applications, 235. Birkhäuser/Springer, Basel, 2014.

[29] Eidelman, Y., Gohberg, I., Olshevsky V., *The QR iteration method for Hermitian quasiseparable matrices of an arbitrary order*, Linear Algebra Appl., 404:305-324, 2005.

[30] Elsner, L., Koltracht, I., Neumann, M., Xiao, D., *On computations of the Perron root*, SIAM J. Matrix Anal. Appl.,14(2):456-467, 1993.

[31] Fasino, D., Gemignani, L., *Structural and computational properties of possibly singular semiseparable matrices*, Linear Algebra Appl., 340:183-198, 2001.

[32] Ferreira, C., Parlett, B., Dopico, F.M., *Sensitivity of eigenvalues of an unsymmetric tridiagonal matrix*, Numer. Math., 122(3):527-555, 2012.

[33] Fiedler, M., *Structure rank of matrices*, Linear Algebra Appl., 179:1199-127, 1993.

[34] Fiedler, M., Markham, T.L., *Completing a matrix when certain entries of its inverse are specified*, Linear Algebra Appl., 74:225-237, 1986.

[35] Fiedler, M., Markham T.L., *Generalized totally nonnegative matrices*, Linear Algebra Appl., 345:9-28, 2002.

[36] Gantmacher, F.R., Krein, M.G., *Oscillation Matrices and Kernels and Small Vibrations of Mechanical Systems*, AMS Chelsea Publishing, Providence, Rhode Island, 2002 (revised edition from the Russian original edition published in 1941).

[37] Geurts, A.J., *A contribution to the theory of condition*, Numer. Math., 39:85-96, 1982.

[38] Golub, G.H., Van Loan, C.F., *Matrix Computations, 3rd ed.*, The Johns Hopkins University Press, Baltimore, 1996.

[39] Grasedyck, L., Kriemann, R., Le Borne, S., *Parallel black box H-LU preconditioning for elliptic boundary value problems*, Comput. Vis. Sci., 11:273-291, 2008.

[40] Gustafson, W.H., *A note on matrix inversion*, Linear Algebra Appl., 57:71-73, 1984.

[41] Higham, N.J., *Accuracy and Stability of Numerical Algorithms, 2nd ed.*, Society for Industrial and Applied Mathematics, Philadelphia, 2002.

[42] Higham, D.J., Higham, N.J., *Structured backward error and condition of generalized eigenvalue problems*, SIAM J. Matrix Anal. Appl., 20(2):493-512, 1998.

[43] Higham, D.J., Higham, N.J., *Backward error and condition of structured linear systems*, SIAM J. Matrix Anal. Appl., 13(1):162-175, 1992.

[44] Horn, R.A., Johnson, C.R., *Topics in Matrix Analysis*, Cambridge University Press, Cambridge, 1991.

[45] Ipsen, I.C.F., *Relative perturbation results for matrix eigenvalues and singular values*, Acta Numerica, 7:151-201, 1998.

[46] Karow M., Kressner D., Tisseur F, *Structured eigenvalue condition numbers*, SIAM J. Matrix Anal. Appl., 28(4):1052-1068, 2006.

[47] Kellogg O. D., *The oscillation of functions of an orthogonal set*, Am. J. Math., 38:1-5, 1916.

[48] Kellogg O. D., *Orthogonal functions sets arising from integral equations*, Am. J. Math., 40:145-154, 1918.

[49] Li, R.C., *Relative perturbation theory. III. More bounds on eigenvalue variations*, Linear Algebra Appl., 266:337-345, 1997.

[50] Martinsson, P.G., Rokhlin V., *A fast direct solver for boundary integral equations in two dimensions*, J. Comput. Phys., 205:1-23, 2005.

[51] Martinsson, P.G., Rokhlin, V., Tygert, M., *A fast algorithm for the inversion of general Toeplitz matrices*, Comput. Math. Appl., 50:741-752, 2005.

[52] Parlett, B.N., Reinsch, C., *Balancing a matrix for calculation of eigenvalues and eigenvectors*, Numer. Math., 13:292-304, 1969.

[53] Rice, J.R., *A theory of condition*, SIAM J. Numer. Anal., 3:287-310, 1966.

[54] Robol, L., *Exploiting rank structures for the numerical treatment of matrix polynomials*, PhD Thesis, Scuola Normale Superiore, Pisa, 2015.

[55] Roy, S.N., Sarhan A.E., *On inverting a class of patterned matrices*, Biometrika, 43:227-231, 1956.

[56] Stewart, G.W., Sun, J., *Matrix Perturbation Theory*, Academic Press, Boston, 1990.

[57] Stewart, M., *On orthogonal transformations of rank structured matrices*, submitted (available at `http://saaz.mathstat.gsu.edu/pub/pub.html`).

[58] Van Barel, M., Vandebril, R., Van Dooren, P., Frederix, K., *Implicit double shift QR-algorithm for companion matrices*, Numer. Math., 116:177-212, 2010.

[59] Vandebril, R., Van Barel, M., Mastronardi, N., *A note on the representation and definition of semiseparable matrices*, Numer. Linear Algebra Appl., 12:839-858, 2005.

[60] Vandebril, R., Van Barel, M., Golub G., Mastronardi, N., *A bibliography on semiseparable matrices*, Calcolo, 42:249-270, 2005.

[61] Vandebril, R., Van Barel, M., Mastronardi, N., *Matrix Computations and Semiseparable Matrices. Volume 1. Linear Systems*, The Johns Hopkins University Press, Baltimore, 2008.

[62] Vandebril, R., Van Barel, M., Mastronardi, N., *Matrix Computations and Semiseparable Matrices. Volume 2. Eigenvalue and Singular value methods*, The Johns Hopkins University Press, Baltimore, 2008.

[63] Van Camp, E., *Diagonal-plus-semiseparable matrices and their use in Numerical Linear Algebra* , PhD Thesis, KU Leuven, 2005.

[64] Van Camp, E., Van Barel M., Vandebril, R., Mastronardi, N., *Two fast algorithms for solving diagonal-plus-semiseparable linear systems*, J. Comput. Appl. Math., 164-165, 2004.

[65] Wilkinson, J.H., *The Algebraic Eigenvalue Problem*, Oxford University Press, New York, 1965.

[66] Xi, Y., Xia, J., *On the stability of some hierarchical rank structured matrix algorithms*, submitted.

[67] Xi, Y., Xia, J., Cauley, S., Balakrishnan, V., *Superfast and stable structured solvers for Toeplitz least squares via randomized sampling*, SIAM J. Matrix Anal. Appl, 35(1):44-72, 2014.