The conditioning of FD matrix sequences coming from semi-elliptic Differential Equations [1]

D. Noutsos [a] S. Serra Capizzano [b] P. Vassalos [c]

[a] *University of Ioannina, Department of Mathematics, T.K 45110 Ioannina, Greece*

[b] *Università dell'Insubria - Sede di Como, Dipartimento di Fisica e Matematica, Via Valleggio 11 - 22100 COMO, Italy.*

[c] *Athens University of Business and Economics, Department of Informatics, Patission 76, T.K 10434, Athens, Greece*

Abstract

In this paper we are concerned with the study of spectral properties of the sequence of matrices $\{A_n(a)\}$ coming from the discretization, using centered finite differences of minimal order, of elliptic (or semielliptic) differential operators $L(a,u)$ of the form

$$\begin{cases} -\frac{d}{dx}\left(a(x)\frac{d}{dx}u(x)\right) = f\,(x) & \text{on } \Omega = (0,1), \\ \text{Dirichlet B.C. on } \partial\Omega, \end{cases} \tag{1}$$

where the nonnegative, bounded coefficient function $a(x)$ of the differential operator may have some isolated zeros in $\overline{\Omega} = \Omega \cup \partial\Omega$. More precisely, we state and prove the explicit form of the inverse of $\{A_n(a)\}$ and some formulas concerning the relations between the orders of zeros of $a(x)$ and the asymptotic behavior of the minimal eigenvalue (condition number) of the related matrices. As a conclusion, and in connection with our theoretical findings, first we extend the analysis to higher order (semi-elliptic) differential operators, and then we present various numerical experiments, showing that similar results must hold true in 2D as well.

Key words: Finite Differences, Toeplitz matrices, Boundary Value Problems, Spectral Distribution.
1991 MSC: 65N22, 65F10, 15A12.

Email addresses: `dnoutsos@uoi.gr` (D. Noutsos), `stefano.serrac@uninsubria.it` (S. Serra Capizzano), `pvassal@aueb.gr` (P. Vassalos).

[1] This research was co-funded by the European Union - European Social Fund (ESF) & National Sources, in the framework of the program "Pythagoras I" of the

1 Introduction

The numerical solution of elliptic 1D and 2D Boundary Value Problems (BVPs) is a classical topic arising from a wide range of applications such as elasticity problems, nuclear and petroleum engineering etc. [31]. In these contexts, the coefficient function can be continuous or discontinuous, but its positivity guarantees the ellipticity of the continuous problem. On the other hand, for the calculation of special functions or for applications to mathematical biology and mathematical finance, the strict ellipticity is lost and indeed the function may have isolated zeros generally located at the boundary $\partial\Omega$ of the definition domain (see [16,32,1] and references therein).

Since the arising linear systems are of large size, fast and efficient resolution methods are always welcome and, for stability reasons, iterative techniques have to be preferred. However, in order to devise efficient and accurate iterative procedures, crucial spectral properties of $\{A_n(a)\}$ must be understood. In particular, we are interested in spectral localization results and especially in the asymptotic behavior of the extreme eigenvalues (which implies the knowledge of the asymptotical conditioning). Furthermore, the characterization and understanding of the subspace where the ill-conditioning occurs would be also useful, at least in a certain approximate sense. In fact the latter information represents a theoretical basis for the construction of effective preconditioners for classical and Krylov based iterative methods or in designing good prolongation/restriction operators for multigrid methods (see [15,30] and references therein). In the specific case of elliptic and semi-elliptic non-necessarily symmetric BVPs and positive definite ill-conditioned non-necessarily Hermitian Toeplitz sequences, this approach has been quite successful, both in sequential and parallel models of computation (see [11,12,21,27,24,26,4])

In this paper, we study the asymptotic conditioning with special attention to the minimal eigenvalue, since it is easy to prove that the maximal eigenvalue is bounded by a pure constant (see e.g. [10,25]). From the viewpoint of the mathematical tools, we widely use three notions of positivity: componentwise positivity (so that the Perron-Frobenius theory [31] can be invoked), positive definiteness (so that the evaluation of the spectral norm, induced by the Euclidean vector norm, is reduced to an eigenvalue analysis i.e. to study of the spectral radius), and operator positivity (so that powerful equivalence results can be applied, see [23]).

For problem (1) and for strictly positive coefficient function $a(x)$, in [10] it has been proved that the Euclidean condition number of $A_n(a)$ grows as n^2. For the degenerate case of $a(x)$ with some isolated zeros, in [21], the second author

"Operational Program for Education and Initial Vocational Training" of the 3rd Community Support Framework of the Hellenic Ministry of Education.

argues that the condition number of the arising sequence $\{A_n\}_n$ is affected by two factors (see also [25] and Subsection 2.2): the order of the differential operator which causes a growth of order n^2 (for second order problems) and the order α of the unique zero of the coefficient $a(x)$ which gives a contribution of order n^α.

The main goal of this paper is to give an explicit formula for the inverse of A_n and an asymptotical study of its condition number, for every nonnegative bounded function $a(x)$, not necessarily regular (see the beginning of Section 3 for the precise hypotheses), and with a unique zero: in particular, we show that the conditioning grows as $n^{\max\{\alpha,2\}}$, up at most to the factor $\log(n)$ only in the case where $\alpha = 2$.

The analysis is then extended to the case of several zeros and to the case of higher order operators: more specifically, when more than one zero is involved the behavior of the conditioning becomes less regular and resonance effects appear, increasing the order of the conditioning; on the other hand, for $2k$th order BVPs, $k \geq 1$, and with a unique zero of order α in the nonnegative coefficient, the quantity $n^{\max\{\alpha,2\}}$ is simply (and naturally) replaced by $n^{\max\{\alpha,2k\}}$. Finally, even though we focus our attention on 1D problems, we should stress that an interesting side-effect of this paper is to provide a theoretical framework which can be exploited to cover the less explored and highly interesting multidimensional case.

The paper is organized as follows: in Section 2 we set the problem in more detail, we set notations, and we report in a organized way some more or less known results from the relevant literature; Section 3 is devoted to give the explicit form for the inverse of the matrix A_n, a fundamental tool for our derivations, while, in Section 4, we determine the asymptotic behavior of the spectral radius of A_n^{-1}, for the second order problem in (1). Section 5 is addressed to the extension of our findings in the case of arbitrary order elliptic BVPs. Furthermore, in Section 6 we discuss the extension of our main theorem in 2D, something which is ascertained numerically in Section 7, where several 1D and 2D numerical experiments are presented and discussed. Section 8 is finally devoted to conclusions and perspectives.

2 Definition of the problem, notations, and preliminary results

Let us consider the second order BVP (1) and its approximation by using centered finite differences, of minimal bandwidth, of precision order two, and of stepsize $h = (n+1)^{-1}$ on the grid-points $x_0 = 0, x_1, x_2, \ldots, x_n, x_{n+1} = 1$. If x_t denotes $\dfrac{t}{n+1}$, $t \in [0, n+1]$, $a_t = a(x_t)$, $f_t = f(x_t)$, and u_t represents an approximation of $u(\cdot)$ at x_t, then the considered numerical scheme leads to the following set of equations

$$-a_{i-\frac{1}{2}}u_{i-1} + (a_{i-\frac{1}{2}} + a_{i+\frac{1}{2}})u_i - a_{i+\frac{1}{2}}u_{i+1} = h^2 f_i, \quad i = 1, 2, \ldots, n.$$

Then, by collecting the above formulae and by taking into account the boundary conditions, we arrive to $n \times n$ linear system whose coefficient matrix $A_n(a)$ shows the form

$$\begin{bmatrix} a_{\frac{1}{2}} + a_{\frac{3}{2}} & -a_{\frac{3}{2}} & & & & \\ -a_{\frac{3}{2}} & a_{\frac{3}{2}} + a_{\frac{5}{2}} & -a_{\frac{5}{2}} & & & \\ & -a_{\frac{5}{2}} & \ddots & \ddots & & \\ & & \ddots & \ddots & \ddots & \\ & & & \ddots & \ddots & -a_{n-\frac{1}{2}} \\ & & & & -a_{n-\frac{1}{2}} & a_{n-\frac{1}{2}} + a_{n+\frac{1}{2}} \end{bmatrix} \tag{2}$$

Let $T_n = A_n(1)$ be the Toeplitz matrix (i.e. constant along diagonals) discretizing problem (1) with $a = 1$, that is the matrix in (2) with $a = 1$. The matrices $A_n(a)$ can be expressed as

$$A_n(a) = \sum_{i=1}^{n+1} a_{i-1/2} Q_n(i), \tag{3}$$

where the matrices $Q_n(i)$ are symmetric nonnegative definite dyads given by

$$Q_n(i) = \begin{array}{c} \begin{array}{cccccc} & & & & & i \end{array} \\ \begin{bmatrix} 0 & \cdots & 0 & 0 & \cdots & 0 \\ \vdots & \ddots & \vdots & \vdots & & \vdots \\ 0 & \cdots & 1 & -1 & \cdots & 0 \\ 0 & \cdots & -1 & 1 & \cdots & 0 \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ 0 & \cdots & 0 & 0 & \cdots & 0 \end{bmatrix} \end{array}, \qquad i = 2, \ldots, n,$$

and $Q_n(1) = e_1 e_1^T$, $Q_n(n+1) = e_n e_n^T$, with e_j, $j = 1, \ldots, n$, representing the jth column of the identity matrix.

Therefore the matrix

$$T_n = \sum_{i=1}^{n+1} Q_n(i), \tag{4}$$

is the sum of all the dyads $Q_n(i)$ and $A_n(a)$ is a weighted sum of the same dyads according to the weights $a_{i-1/2}$, $i = 1, 2, \ldots, n+1$. Moreover each dyad has a "local structure" with respect to the canonical basis of $\mathbb{R}^{n\times n}$ so that each weight $a_{i-1/2}$ contributes in the matrix $A_n(a)$ to $E_{i-1,i-1}$, $E_{i,i-1}$, $E_{i-1,i}, E_{i,i}$ where $E_{s,t} = e_s e_t^T$. Furthermore, this notion of "locality" is geometrical as well, since vectors of the canonical basis that are close (e_s and e_t are close if $|s-t|/n = o(1)$) correspond to dyads

$$Q_n(s-1), \quad Q_n(s), \quad Q_n(t-1), \quad Q_n(t)$$

such that the related weights come from close points in the interval $[0,1]$. Therefore we can say that the matrices $\{A_n(a)\}_n$ have a local decomposition with respect to the Toeplitz matrices $\{T_n = A_n(1)\}_n$: this locality principle is important for obtaining global distribution results for the spectra of the related matrix sequences (see e.g. [29,22]). However, again thanks to (3) and to the nonnegative definiteness of the basic dyads $Q_n(i)$, an other important aspect is that $A_n(\cdot)$, regarded as an operator from a suitable function space $\mathcal{S}$ into $\mathbb{R}^{n\times n}$, is linear and positive i.e. $A_n(\alpha a+\beta b) = \alpha A_n(a)+\beta A_n(b)$, $\alpha, \beta \in \mathbb{R}$, $a, b \in \mathcal{S}$ and $A_n(a)$ is nonnegative definite if a is nonnegative, as a function in $\mathcal{S}$ (see [23,28] for a general discussion and several results on matrix-valued linear positive operators). In Subsection 2.2 , we will use (3), (4), and this notion of operator positivity for obtaining preliminary results on the eigenvalues of $A_n(a)$.

Finally we should emphasize that the latter dyadic decompositions have a much broader interest and, in actuality, they apply to general differential operators approximated by general finite differences (see [25, Theorem 4.1] and also Lemma 2.1, Corollary 3.3, and Theorem 3.5 in the same paper) and by finite elements (see Sections 3 and 4 in [3]).

2.1 Notations

We introduce symbols that we will use throughout the paper. Let us consider two nonnegative functions $\alpha(\cdot)$ and $\beta(\cdot)$ defined over a domain D with accumulation point $\bar{x}$ (if $D = \mathbb{N}$ then $\bar{x} = \infty$, if $D = [0,1]^d$, $d = 1, 2$, then $\bar{x}$ can be any point of D). We write

- $\alpha(\cdot) = O(\beta(\cdot))$ if and only if there exists a pure positive constant K such that $\alpha(x) \le K\beta(x)$, for every (or for almost every) $x \in D$ (here and in the following for pure or universal constant we mean a quantity not depending on the variable $x \in D$);
- $\alpha(\cdot) = \Omega(\beta(\cdot))$ if and only if there exists a pure positive constant K such that $\alpha(x) \ge K\beta(x)$, for every (or for almost every) $x \in D$;

- $\alpha(\cdot) = o(\beta(\cdot))$ if and only if $\alpha(\cdot) = O(\beta(\cdot))$ and $\lim_{x \to \bar{x}} \alpha(x)/\beta(x) = 0$ with $\bar{x}$ given accumulation point of D which will be clear from the context;
- $\alpha(\cdot) \sim \beta(\cdot)$ if and only if $\alpha(\cdot) = O(\beta(\cdot))$ and $\beta(\cdot) = O(\alpha(\cdot))$ (or, equivalently, if and only if $\alpha(\cdot) = O(\beta(\cdot))$ and $\alpha(\cdot) = \Omega(\beta(\cdot))$;
- $\alpha(\cdot) \approx \beta(\cdot)$ if and only if $\alpha(\cdot) \sim \beta(\cdot)$ and $\lim_{x \to \bar{x}} \alpha(x)/\beta(x) = 1$ with $\bar{x}$ given accumulation point of D (the latter can be rewritten as $\alpha(x) = \beta(x)(1+o(1))$ with $1 + o(1)$ uniformly positive in D).

2.2 Preliminary results

In the following, with respect to problem (1) and hence with respect to the matrix structure in (2), we assume that the functional coefficient $a(x)$ is bounded, piece-wise continuous, nonnegative, and with a unique zero at 0 of order α i.e. $a(x) \sim x^\alpha$ on $D = [0, 1]$.

Since $A_n(\cdot)$ can be regarded as a matrix-valued linear positive operator, it is clear that it is also monotone (see [23]) that is $A_n(b) \geq A_n(a)$ if $b \geq a$ where, as usual, the ordering is the partial ordering in the space of symmetric real matrices and that of the function space $\mathcal{S}$, respectively. Therefore, since in our context the coefficient $a(x)$ is nonnegative and bounded, it follows that $A_n(a) \leq \|a\|_\infty A_n(1) = \|a\|_\infty T_n$. From the latter, from the monotonicity of the eigenvalues (i.e. $A \leq B$ implies $\lambda_j(A) \leq \lambda_j(B)$, for every pair of $n \times n$ Hermitian matrices and for every index $j = 1, 2, \ldots, n$, where $\lambda_1(X) \leq \lambda_2(X) \leq \cdots \leq \lambda_n(X)$, $X \in \{A, B\}$, see [6]) and from the known expression of the eigenvalues of T_n, we deduce that

$$\lambda_{\min}(A_n) \leq \|a\|_\infty 4 \sin^2\left(\frac{\pi}{2(n+1)}\right) \sim n^{-2}. \tag{5}$$

On the other hand, if $a(x)$ has a unique zero at zero of order α, then the minimal eigenvalue of $A_n = A_n(a)$ tends to zero at least as $n^{-\alpha}$ (see also [21, Proof of Theorem 4.1]). In fact, from (2) and from the Courant-Fisher characterization (see e.g. [6]), we have

$$\lambda_{\min}(A_n) \leq \frac{e_1^T A_n e_1}{e_1^T e_1} = a_{\frac{1}{2}} + a_{\frac{3}{2}} \sim n^{-\alpha}. \tag{6}$$

Therefore the latter bounds imply

$$\lambda_{\min}(A_n) \leq C n^{-\max\{\alpha, 2\}}, \tag{7}$$

with C universal constant independent of n (indeed depending only on the coefficient $a(x)$, see (6)). Conversely, by exploiting again the monotonicity of the operator $A_n(\cdot)$ and of the eigenvalues, and by using the dyadic decomposition in (3), it follows that

$$\lambda_{\min}(A_n) \geq \min_{1\leq i\leq n+1} a_{i-1/2}\lambda_{\min}(T_n) \sim n^{-(\alpha+2)}. \tag{8}$$

Here we are interested in filling the gap between (7) and (8) and in fact, in Section 4, we will prove via Perron-Frobenius tools (see e.g. [31]) that the order of the true behavior of the minimal eigenvalue is described by $n^{-\max\{\alpha,2\}}$, with, at most, an additional factor $\log(n)$ in the case where $\alpha = 2$: that factor could be motivated as a kind of resonance typical of finite differences in presence of multiple zeros in the characteristic polynomial. The latter statement has also important implications concerning eigenvectors: indeed the two sources of ill-conditioning, the low frequencies coming from the constant coefficient Laplacian, and the space spanned by few canonical vectors related to the position of the zero of $a(x)$, do not interfere. There is only a superposition effect so that the size of the degenerating subspace (i.e. that related to small eigenvalues) becomes larger, but the order of ill-conditioning is not worse than that of the two factors separately. Therefore, both for designing multigrid methods or preconditioners, we can treat the two ill-conditioned spaces separately in a multi-iterative sense [20], as already done e.g. in [21] by considering a multiplicative diagonal plus Toeplitz preconditioner: more precisely, the diagonal part takes care of the ill-conditioning induced by the zero of $a(x)$ and the Toeplitz part takes care of that induced by the Laplacian (a similar idea is adapted in [26] in a multigrid setting). Finally we just mention that other results of this type can be found in [21, Theorem 4.1] and [25, Corollary 4.1 and the third item of Theorem 4.3].

3 Explicit form for the inverse of the matrix A_n

Let us consider the second order BVP (1) discretized as described in Section 2. We assume that the functional coefficient $a(x)$ is bounded, piece-wise continuous, nonnegative, and with a unique zero at 0 of order α i.e. $a(x) \sim x^\alpha$ on $D = [0,1]$. The matrix coming from the considered approximation is $A_n = A_n(a)$ as displayed in (2). In the quoted literature, we find several contributions discussing the form of the inverse of a tridiagonal matrix, or more generally, on the one of a band matrix. First in 1960, F. Gantmacher and M. Krein [13] proved that the inverse of a symmetric nonsingular tridiagonal matrix is a Green matrix which is defined by the Hadamard product of a weak type D and a flipped weak type D matrices as follows:

$$C = U \circ V = \begin{bmatrix} u_1 & u_1 & \cdots & u_1 \\ u_1 & u_2 & \cdots & u_2 \\ \vdots & \vdots & \ddots & \vdots \\ u_1 & u_2 & \cdots & u_n \end{bmatrix} \circ \begin{bmatrix} v_1 & v_2 & \cdots & v_n \\ v_2 & v_2 & \cdots & v_n \\ \vdots & \vdots & \ddots & \vdots \\ v_n & v_n & \cdots & v_n \end{bmatrix} = \begin{bmatrix} u_1v_1 & u_1v_2 & \cdots & u_1v_n \\ u_1v_2 & u_2v_2 & \cdots & u_2v_n \\ \vdots & \vdots & \ddots & \vdots \\ u_1v_n & u_2v_n & \cdots & u_nv_n \end{bmatrix}. \quad (9)$$

Conversely, the same authors have proven that the inverse of a Green matrix is a symmetric tridiagonal matrix. In 1970, M. Capovani [8] stated and derived relations which give the entries of the inverse of a tridiagonal matrix in terms of its entries and its subdeterminants. In the same paper he gave the form of the inverse of some particular cases of tridiagonal and block tridiagonal matrices. One year later the same author [9], extended the result of F. Gantmacher and M. Krein [13] for nonsymmetric matrices. R. Bevilacqua and M. Capovani [5] in 1976, gave structural properties to determine the coefficients of the inverse of a (block) as a function of its (blocks) entries. In 1979, W. Barrett [2] proved that a matrix R with $R_{22}, \ldots, R_{n-1,n-1} \neq 0$ has the triangle property if and only if its inverse is a tridiagonal matrix: more in detail, a matrix R has this useful property if $R_{ij} = \frac{R_{ik}R_{kj}}{R_{kk}}$ for all $i < k < j$ and all $i > k > j$. In 1987, P. Rózsa [19], using properties of Green's matrices and of semi-separable matrices, proposed an algorithm to determine the elements of the inverse of a band matrix by solving some difference equations. Later in 1998, J. McDonald, R. Nabben, M. Neumann, H. Schneider and M. Tsatsomeros [17] generalized the result of F. Gantmacher and M. Krein [13] for nonsymmetric tridiagonal Z-matrices and they proved properties for the inverse of a tridiagonal M-matrix. They gave also properties for the inverse of such matrices in terms of special structured matrices called cyclopses (see again [17] for a formal definition). More recently, i.e. in 1999, R. Nabben [18] proved properties for the inverse of tridiagonal M, positive definite and diagonally dominant matrices.

The matrix A_n in (2) has most of the above "good" properties: it is an irreducible nonsingular tridiagonal Z-matrix, an M-matrix, and also a symmetric positive definite matrix. Hence, we can combine the above results for characterizing its inverse. However, the matrix A_n has an additional property that all row sums are zeros except the first and the last one. Taking into account Corollary 3.6 of [17] or Corollary 2.6 of [18], concerning properties of the inverse of an M-matrix, and Corollary 2.7 of [18], concerning on properties of the inverse of a positive definite matrix, we obtain that the numbers $u_i, v_i,\ i = 1, 2, \ldots, n$, appearing in the Hadamard product (9), can be chosen to be positive and such that

$$0 < \frac{u_1}{v_1} < \frac{u_2}{v_2} < \cdots < \frac{u_n}{v_n}. \quad (10)$$

In the sequel we will find an explicit form for the matrix A_n^{-1} by using the forms of A_n and C in (2) and (9), respectively and inequalities (10). We take the product A_nC which should be the identity matrix I.

For $k < j$, the inner product of the kth row of A_n with the jth column of C gives

$$\begin{aligned} 0 = (A_nC)_{kj} &= v_j \left(-a_{k-\frac{1}{2}} u_{k-1} + (a_{k-\frac{1}{2}} + a_{k+\frac{1}{2}}) u_k - a_{k+\frac{1}{2}} u_{k+1} \right) \\ &= v_j \left(a_{k-\frac{1}{2}} (u_k - u_{k-1}) - a_{k+\frac{1}{2}} (u_{k+1} - u_k) \right). \end{aligned}$$

We observe that this equality holds true if we chose, up to a constant factor,

$$u_k - u_{k-1} = \frac{1}{a_{k-\frac{1}{2}}}, \quad k = 2, 3, \ldots, n.$$

One solution of this difference equation, up to a constant factor, is

$$u_k = \sum_{i=1}^{k} \frac{1}{a_{i-\frac{1}{2}}}, \quad k = 1, 2, 3, \ldots, n.$$

For $k = 1$ we have

$$0 = (A_nC)_{1j} = v_j \left((a_{\frac{1}{2}} + a_{\frac{3}{2}}) \frac{1}{a_{\frac{1}{2}}} - a_{\frac{3}{2}} \left(\frac{1}{a_{\frac{1}{2}}} + \frac{1}{a_{\frac{3}{2}}} \right) \right)$$

which holds true.

For $k > j$, the associated inner products give

$$\begin{aligned} 0 = (A_nC)_{kj} &= u_j \left(-a_{k-\frac{1}{2}} v_{k-1} + (a_{k-\frac{1}{2}} + a_{k+\frac{1}{2}}) v_k - a_{k+\frac{1}{2}} v_{k+1} \right) \\ &= u_j \left(-a_{k-\frac{1}{2}} (v_{k-1} - v_k) + a_{k+\frac{1}{2}} (v_k - v_{k+1}) \right). \end{aligned}$$

We observe also here that we can chose, up to a constant factor,

$$v_{k-1} - v_k = \frac{1}{a_{k-\frac{1}{2}}}, \quad k = 2, 3, \ldots, n.$$

One solution of this difference equation, up to a constant factor, is

$$v_k = \sum_{i=k}^{n} \frac{1}{a_{i+\frac{1}{2}}}, \quad k = 1, 2, 3, \ldots, n.$$

For $k = n$ we have

$$0 = (A_nC)_{nj} = u_j \left(-a_{n-\frac{1}{2}} \left(\frac{1}{a_{n-\frac{1}{2}}} + \frac{1}{a_{n+\frac{1}{2}}} \right) + (a_{n-\frac{1}{2}} + a_{n+\frac{1}{2}}) \frac{1}{a_{n+\frac{1}{2}}} \right)$$

which holds true. We define by s_k and by s the sums $\sum_{i=k}^{n} \frac{1}{a_{i+\frac{1}{2}}}$ and $\sum_{i=0}^{n} \frac{1}{a_{i+\frac{1}{2}}}$, respectively. It is obvious that with the above choices, up to a constant factor, we have $v_k = s_k, \quad u_k = s - s_k, \quad k = 1, 2, \ldots, n$. We observe also that the sequence v_k strictly decreases while u_k strictly increases, so the inequalities (10) are satisfied.

It remains to check the inner products for $k = j$.

$$\begin{aligned}
(A_nC)_{kk} &= -a_{k-\frac{1}{2}}u_{k-1}v_k + (a_{k-\frac{1}{2}} + a_{k+\frac{1}{2}})u_kv_k - a_{k+\frac{1}{2}}u_kv_{k+1} \\
&= -a_{k-\frac{1}{2}}(s - s_{k-1})s_k + (a_{k-\frac{1}{2}} + a_{k+\frac{1}{2}})(s - s_k)s_k - a_{k+\frac{1}{2}}(s - s_k)s_{k+1} \\
&= a_{k-\frac{1}{2}}(s_{k-1} - s_k)s_k + a_{k+\frac{1}{2}}(s_k - s_{k+1})(s - s_k) \\
&= s_k + (s - s_k) = s, \quad k = 2, 3, \ldots, n-1,
\end{aligned}$$

$$\begin{aligned}
(A_nC)_{11} &= (a_{\frac{1}{2}} + a_{\frac{3}{2}})u_1v_1 - a_{\frac{3}{2}}u_1v_2 = a_{\frac{1}{2}}(s - s_1)s_1 + a_{\frac{3}{2}}(s_1 - s_2)(s - s_1) \\
&= s_1 + (s - s_1) = s,
\end{aligned}$$

$$\begin{aligned}
(A_nC)_{nn} &= -a_{n-\frac{1}{2}}u_{n-1}v_n + (a_{n-\frac{1}{2}} + a_{n+\frac{1}{2}})u_nv_n \\
&= a_{n-\frac{1}{2}}(s_{n-1} - s_n)s_n + a_{n+\frac{1}{2}}(s - s_n)s_n = s_n + (s - s_n) = s.
\end{aligned}$$

As a consequence $A_nC = sI$. To eliminate s we have to chose the constant factors of the matrices U and V, in such a way that the relative product equals $\frac{1}{s}$. Then, the inverse of A_n is obtained by dividing C by s which gives us the explicit form:

$$A_n^{-1} = \begin{bmatrix}
\frac{s_1(s-s_1)}{s} & \frac{s_2(s-s_1)}{s} & \frac{s_3(s-s_1)}{s} & \ldots & \frac{s_n(s-s_1)}{s} \\
\frac{s_2(s-s_1)}{s} & \frac{s_2(s-s_2)}{s} & \frac{s_3(s-s_2)}{s} & \ldots & \frac{s_n(s-s_2)}{s} \\
\frac{s_3(s-s_1)}{s} & \frac{s_3(s-s_2)}{s} & \frac{s_3(s-s_3)}{s} & \ldots & \frac{s_n(s-s_3)}{s} \\
\vdots & \vdots & \vdots & \ddots & \vdots \\
\frac{s_n(s-s_1)}{s} & \frac{s_n(s-s_2)}{s} & \frac{s_n(s-s_3)}{s} & \ldots & \frac{s_n(s-s_n)}{s}
\end{bmatrix}. \tag{11}$$

It follows another proof to obtain the explicit form of A_n^{-1} independent of the previous approach. It is not related to the theory appearing in the referred literature, but only depends on the form (2) of A_n and on a tricky use of the Sherman-Morrison formula.

We consider the matrix

$$\tilde{A}_n = \text{tridiag}[-1 \ \ 1 \ \ 0]\text{diag}[a_{\frac{3}{2}} \ \ a_{\frac{5}{2}} \ \ \cdots \ \ a_{n+\frac{1}{2}}]\text{tridiag}[0 \ \ 1 \ \ -1]$$

i.e.

$$\begin{aligned}\tilde{A}_n &= \begin{bmatrix} 1 & & & & \\ -1 & 1 & & & \\ 0 & -1 & 1 & & \\ \vdots & \vdots & \ddots & \ddots & \\ 0 & 0 & \cdots & -1 & 1 \end{bmatrix} \begin{bmatrix} a_{\frac{3}{2}} & & & & \\ & a_{\frac{5}{2}} & & & \\ & & a_{\frac{7}{2}} & & \\ & & & \ddots & \\ & & & & a_{n+\frac{1}{2}} \end{bmatrix} \begin{bmatrix} 1 & -1 & 0 & \cdots & 0 \\ & 1 & -1 & \cdots & 0 \\ & & 1 & \ddots & \vdots \\ & & & \ddots & -1 \\ & & & & 1 \end{bmatrix} \\ &= \begin{bmatrix} a_{\frac{3}{2}} & -a_{\frac{3}{2}} & & & & \\ -a_{\frac{3}{2}} & a_{\frac{3}{2}} + a_{\frac{5}{2}} & -a_{\frac{5}{2}} & & & \\ & -a_{\frac{5}{2}} & \ddots & \ddots & & \\ & & \ddots & \ddots & \ddots & \\ & & & \ddots & \ddots & -a_{n-\frac{1}{2}} \\ & & & & -a_{n-\frac{1}{2}} & a_{n-\frac{1}{2}} + a_{n+\frac{1}{2}} \end{bmatrix}.\end{aligned}$$

We observe that the matrices $\tilde{A}_n$ and A_n differ only in position $(1,1)$, where the term $a_{\frac{1}{2}}$ does not appear in the matrix $\tilde{A}_n$. By (3), it follows that

$$A_n = \tilde{A}_n + a_{\frac{1}{2}} e_1 e_1^T,$$

e_1 being the first column of the identity matrix. Our aim is to find an explicit form of the inverse of A_n and hence, from the above relation, we write

$$A_n^{-1} = (\tilde{A}_n + a_{\frac{1}{2}} e_1 e_1^T)^{-1} = (I + a_{\frac{1}{2}} \tilde{A}_n^{-1} e_1 e_1^T)^{-1} \tilde{A}_n^{-1}. \tag{12}$$

As a consequence, in order to determine a formula for the inverse of A_n, it suffices to compute the inverses of the two factors appearing in (12). For that purpose, we start our analysis by studying the inverse of $\tilde{A}_n$ (in this respect, we should acknowledge that the expression of $\tilde{A}_n^{-1}$ can be found, for the specific case of $a = 1$, in [7, Chapter 4, Exercise 8, p. 108]).

From the above factorization of A_n, we find

$$\tilde{A}_n^{-1} = (\mathrm{tridiag}[0 \quad 1 \quad -1])^{-1} (\mathrm{diag}[a_{\frac{3}{2}} \quad a_{\frac{5}{2}} \quad \cdots \quad a_{n+\frac{1}{2}}])^{-1} (\mathrm{tridiag}[-1 \quad 1 \quad 0])^{-1}$$

that is

$$\tilde{A}_n^{-1} = \begin{bmatrix} 1\,1\,1 & \cdots & 1 \\ & 1\,1\cdots & 1 \\ & 1 \ddots & \vdots \\ & \ddots & 1 \\ & & 1 \end{bmatrix} \begin{bmatrix} \frac{1}{a_{\frac{3}{2}}} & & & & \\ & \frac{1}{a_{\frac{5}{2}}} & & & \\ & & \frac{1}{a_{\frac{7}{2}}} & & \\ & & & \ddots & \\ & & & & \frac{1}{a_{n+\frac{1}{2}}} \end{bmatrix} \begin{bmatrix} 1 & & & & \\ 1 & 1 & & & \\ 1 & 1 & 1 & & \\ \vdots & \vdots & \ddots & \ddots & \\ 1 & 1 & \cdots & 1 & 1 \end{bmatrix}$$

$$= \begin{bmatrix} \frac{1}{a_{\frac{3}{2}}} & \frac{1}{a_{\frac{5}{2}}} & \frac{1}{a_{\frac{7}{2}}} & \cdots & \frac{1}{a_{n+\frac{1}{2}}} \\ & \frac{1}{a_{\frac{5}{2}}} & \frac{1}{a_{\frac{7}{2}}} & \cdots & \frac{1}{a_{n+\frac{1}{2}}} \\ & & \frac{1}{a_{\frac{7}{2}}} & \ddots & \vdots \\ & & & \ddots & \frac{1}{a_{n+\frac{1}{2}}} \\ & & & & \frac{1}{a_{n+\frac{1}{2}}} \end{bmatrix} \begin{bmatrix} 1 & & & & \\ 1 & 1 & & & \\ 1 & 1 & 1 & & \\ \vdots & \vdots & \ddots & \ddots & \\ 1 & 1 & \cdots & 1 & 1 \end{bmatrix}$$

$$= \begin{bmatrix} \sum_{i=1}^n \frac{1}{a_{i+\frac{1}{2}}} & \sum_{i=2}^n \frac{1}{a_{i+\frac{1}{2}}} & \sum_{i=3}^n \frac{1}{a_{i+\frac{1}{2}}} & \cdots & \frac{1}{a_{n+\frac{1}{2}}} \\ \sum_{i=2}^n \frac{1}{a_{i+\frac{1}{2}}} & \sum_{i=2}^n \frac{1}{a_{i+\frac{1}{2}}} & \sum_{i=3}^n \frac{1}{a_{i+\frac{1}{2}}} & \cdots & \frac{1}{a_{n+\frac{1}{2}}} \\ \sum_{i=3}^n \frac{1}{a_{i+\frac{1}{2}}} & \sum_{i=3}^n \frac{1}{a_{i+\frac{1}{2}}} & \sum_{i=3}^n \frac{1}{a_{i+\frac{1}{2}}} & \cdots & \frac{1}{a_{n+\frac{1}{2}}} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \frac{1}{a_{n+\frac{1}{2}}} & \frac{1}{a_{n+\frac{1}{2}}} & \frac{1}{a_{n+\frac{1}{2}}} & \cdots & \frac{1}{a_{n+\frac{1}{2}}} \end{bmatrix}$$

$$= \begin{bmatrix} s_1 & s_2 & s_3 & \cdots & s_n \\ s_2 & s_2 & s_3 & \cdots & s_n \\ s_3 & s_3 & s_3 & \cdots & s_n \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ s_n & s_n & s_n & \cdots & s_n \end{bmatrix}.$$

The matrix $a_{\frac{1}{2}}\tilde{A}_n^{-1}e_1e_1^T$ has nonzero elements only in the first column whose expression is given by $a_{\frac{1}{2}}s_k$, $k = 1, 2, \ldots, n$. Therefore the analysis of A_n^{-1} is equivalently transformed into the inversion of the matrix

$$I + a_{\frac{1}{2}} \tilde{A}_n^{-1} e_1 e_1^T = \begin{bmatrix} a_{\frac{1}{2}} s & & & & \\ a_{\frac{1}{2}} s_2 & 1 & & & \\ a_{\frac{1}{2}} s_3 & & 1 & & \\ \vdots & & & \ddots & \\ a_{\frac{1}{2}} s_n & & & & 1 \end{bmatrix},$$

where the entry in position $(1,1)$ has been obtained by $1 + a_{\frac{1}{2}} s_1 = a_{\frac{1}{2}} (\frac{1}{a_{\frac{1}{2}}} + s_1) = a_{\frac{1}{2}} s$. It is well-known that the inverse of the above matrix maintains the same structure (since it is a slight variation of an elementary Gauss matrix, see [14]), and it is easily obtained as

$$(I + a_{\frac{1}{2}} \tilde{A}_n^{-1} e_1 e_1^T)^{-1} = \begin{bmatrix} \frac{1}{a_{\frac{1}{2}}} \frac{1}{s} & & & & \\ -\frac{s_2}{s} & 1 & & & \\ -\frac{s_3}{s} & & 1 & & \\ \vdots & & & \ddots & \\ -\frac{s_n}{s} & & & & 1 \end{bmatrix}.$$

In conclusion

$$A_n^{-1} = \begin{bmatrix} \frac{s-s_1}{s} & & & & \\ -\frac{s_2}{s} & 1 & & & \\ -\frac{s_3}{s} & & 1 & & \\ \vdots & & & \ddots & \\ -\frac{s_n}{s} & & & & 1 \end{bmatrix} \begin{bmatrix} s_1 & s_2 & s_3 & \cdots & s_n \\ s_2 & s_2 & s_3 & \cdots & s_n \\ s_3 & s_3 & s_3 & \cdots & s_n \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ s_n & s_n & s_n & \cdots & s_n \end{bmatrix},$$

where we have replaced $\frac{1}{a_{\frac{1}{2}}}$ by $s - s_1$. The above matrix product leads to the following results: the entries of the first row of A_n^{-1} are

$$(A_n^{-1})_{1j} = \frac{s_j(s - s_1)}{s}, \quad j = 1, 2, \ldots, n.$$

The entries $(A_n^{-1})_{ij}$, $i = 2, 3, \ldots, n$ for $i \geq j$, are given by

$$(A_n^{-1})_{ij} = -\frac{s_i s_j}{s} + s_i = \frac{s_i(s - s_j)}{s},$$

while, for $i < j$, we find

$$(A_n^{-1})_{ij} = -\frac{s_i s_j}{s} + s_j = \frac{s_j(s - s_i)}{s}.$$

Consequently a compact formula for the inverse of the matrix A_n is given by the explicit form (11)

4 The spectral radius of A_n^{-1}

For determining the asymptotic behavior of the condition number of the matrix A_n, we have to estimate the smallest eigenvalue since its maximal eigenvalue is bounded by $4\|a\|_\infty$, since $A_n = A_n(a) \leq \|a\|_\infty T_n$ by operator positivity of $A_n(\cdot)$ (see Subsection 2.2) and since $\lambda_{\max}(T_n) < 4$ by Gershgorin's theorem (see e.g. [6,31]). Instead of this, we study the spectral radius of the inverse of A_n. The matrix A_n^{-1} is a symmetric positive definite matrix with positive elements. Thus we make use of the Perron Frobenius theory (see e.g. [31]) for positive (nonnegative) matrices. Our analysis is obtained via a series of preliminary results.

Lemma 4.1. *Let $\{A_n\}_n$, $A_n \in \mathbb{R}^{n\times n}$, be a sequence of symmetric positive definite, irreducible, and nonnegative matrices. If there exists a number $g(n)$ of rows such that their row sums are greater than or equal to $f(n)$, then the order of the spectral radius $\rho(A_n)$ is greater than or equal to $\frac{g(n)f(n)}{n}$, so that*

$$\rho(A_n) = \Omega\left(\frac{g(n)f(n)}{n}\right).$$

Proof. Without loss of generality, we suppose that the row sums are in decreasing order (otherwise this can be obtained by a proper permutation similarity transformation). By using the Courant-Fisher characterization [6], we find

$$\begin{aligned}\rho(A_n) &= \sup_{x\in\mathbb{R}^n, \|x\|=1} x^T A_n x \geq \frac{1}{n} e^T(n) A_n e(n) \\ &= \frac{1}{n} e^T(n) \begin{pmatrix} \tilde{S}_1 \\ \tilde{S}_2 \\ \vdots \\ \tilde{S}_n \end{pmatrix} \geq \frac{1}{n} \sum_{i=1}^{g(n)} \tilde{S}_i \geq \frac{1}{n} g(n) f(n),\end{aligned}$$

where the normalized vector $\frac{1}{\sqrt{n}} e(n)$ has replaced x, with $e(n)$ being the vector of all ones, and where we have denoted by $\tilde{S}_i$ the ith row sum of the matrix A_n. □

We introduce the following definition.

Definition 4.1. *A symmetric, positive definite, irreducible, and nonnegative matrix $A \in \mathbb{R}^{n \times n}$, given in decreasing order of row sums, is dominated by the first $g(n) \times g(n)$ block if $\tilde{S}_i \sim \tilde{S}_{B_i}$, where $\tilde{S}_{B_i} = \sum_{j=1}^{g(n)} a_{ij}$ and the symbol $\sim$ is that defined in Subsection 2.1.*

Lemma 4.2. *Let $\{A_n\}_n$, $A_n \in \mathbb{R}^{n \times n}$, be a sequence of symmetric positive definite, nonnegative matrices, which are dominated by their first $g(n) \times g(n)$ block. If $f(n)$ is the smallest row sum of the first $g(n)$ rows, then the order of the spectral radius $\rho(A_n)$ is greater than or equal to $f(n)$, so that*

$$\rho(A_n) = \Omega(f(n)).$$

Proof. The proof follows the same procedure as of in Lemma 4.1. For that we take the normalized vector $\frac{1}{\sqrt{g(n)}} e(g(n))$, with $e(g(n))$ being the vector of ones in the first $g(n)$ entries and zeros otherwise. Thus

$$\begin{aligned}
\rho(A_n) &= \sup_{x \in \mathbb{R}^n, \|x\|=1} x^T A_n x \\
&\geq \tfrac{1}{g(n)} e^T(g(n)) A_n e(g(n)) \\
&\simeq \tfrac{1}{g(n)} e^T(g(n)) \begin{pmatrix} \tilde{S}_1 \\ \tilde{S}_2 \\ \vdots \\ \tilde{S}_{g(n)} \\ \vdots \end{pmatrix} \\
&= \tfrac{1}{g(n)} \sum_{i=1}^{g(n)} \tilde{S}_i \geq \tfrac{g(n) f(n)}{g(n)} = f(n)
\end{aligned}$$

and the proof is complete. □

Lemma 4.3. *Let $\{A_n\}_n$, $A_n \in \mathbb{R}^{n \times n}$, be a sequence of symmetric positive definite, nonnegative matrices, which are dominated by their first $g(n) \times g(n)$ block. If all the first $g(n)$ rows are of the same order of $f(n)$, then the spectral radius $\rho(A_n)$ is exactly of order $f(n)$.*

Proof. From Lemma 4.2 we deduce $\rho(A_n) = \Omega(f(n))$. On the other hand, from the Perron-Frobenius theory, we obtain that $\rho(A_n) \leq \max_i \tilde{S}_i$. As a consequence, $\rho(A_n) = O(f(n))$ and the proof is complete. □

Now, we are ready to state and prove the main theorem of this section concerning the relation between the order of the zero of the coefficient function $a(x)$ and the condition number of the matrix A_n.

Theorem 4.4. *Let $\{A_n\}_n$, $A_n \in \mathbb{R}^{n \times n}$, be the sequence of matrices derived from the discretization of the Semielliptic Differential Equation (1) with the* *bounded coefficient function $a(x)$ having a unique root at 0 of order α i.e. $a(x) \sim x^\alpha$ on $D = [0,1]$. Then, for the spectral condition number $\kappa(A_n)$ of the matrix A_n which coincides in order with the spectral radius of A_n^{-1}, we find*

$$\kappa(A_n) \sim \rho(A_n^{-1}) \sim \begin{cases} n^2, & 0 \leq \alpha < 2, \\ O(n^2 \log(n)) \cap \Omega(n^2), & \alpha = 2, \\ n^\alpha, & \alpha > 2. \end{cases} \tag{13}$$

Proof. The part $\kappa(A_n) \sim \rho(A_n^{-1})$ simply follows from the relations $\|A_n^{-1}\| = \rho(A_n^{-1})$, $\|A_n\| = \rho(A_n) \leq 4\|a\|_\infty$, and $\lim_{n\to\infty} \rho(A_n) = 4\|a\|_\infty$, where the positive definiteness of A_n and the distribution results in [29] come into the play.

The fact that $a(x) \sim x^\alpha$ means that there exist positive constants c and C far from zero and infinity such that, uniformly with respect to $x \in [0,1]$, we have

$$cx^\alpha \leq a(x) \leq Cx^\alpha.$$

From the positivity of the operator $A_n(\cdot)$ we obtain

$$cA_n(x^\alpha) \leq A_n(a(x)) \leq CA_n(x^\alpha)$$

where the meaning of the inequalities is in the sense of the partial ordering in the real space of Hermitian (real symmetric) matrices. The latter implies

$$c\lambda_i(A_n(x^\alpha)) \leq \lambda_i(A_n(a(x))) \leq C\lambda_i(A_n(x^\alpha)), \quad i = 1, 2, \ldots, n,$$

and, in particular, this holds also for the minimal eigenvalue, which means that the minimal eigenvalue of $A_n(x^\alpha)$ and the minimal eigenvalue of $A_n(a(x))$ coincide in order of magnitude. As a consequence, it is enough to reduce our study to the matrix $A_n(x^\alpha)$, instead of $A_n(a(x))$.

For the remaining part, since the matrix A_n^{-1} is a symmetric positive definite matrix with positive elements, we will prove our assertion, by estimating the row sums of the matrix A_n^{-1} with functional coefficient x^α, given in its explicit form (11), and by using the previous lemmas. For this we study the following cases:

Case 1: $\alpha = 0$.

The result related to this case is well-known [10], since the matrix T_n coincides exactly with tridiag$[-1 \quad 2 \quad -1]$, i.e. the Laplace matrix with eigenvalues $4\sin^2\left(\dfrac{j\pi}{2(n+1)}\right)$, $j = 1, \ldots, n$. Hence

$$\kappa(A_n) \sim \rho(A_n^{-1}) \sim n^2. \tag{14}$$

We remark here that this result could be obtained also by following the reasoning we will use in the subsequent cases.

Case 2: $0 < \alpha < 1$.

We estimate the kth row sum $\tilde{S}_k$ of A_n^{-1}

$$\tilde{S}_k = \frac{S_k}{S}\sum_{i=1}^{k}(S - S_i) + \frac{S - S_k}{S}\sum_{i=k+1}^{n} S_i. \tag{15}$$

First we consider the quantity S_i:

$$\begin{aligned} S_i &= \sum_{j=i}^{n} \frac{1}{a_{j+\frac{1}{2}}} = \sum_{j=i}^{n} \frac{1}{\left(\frac{j+\frac{1}{2}}{n+1}\right)^{\alpha}} = \sum_{j=i}^{n}\left(\frac{2j+1}{2(n+1)}\right)^{-\alpha} \\ &= (n+1)\sum_{j=i}^{n}\left(\frac{2j+1}{2(n+1)}\right)^{-\alpha}\frac{1}{n+1}. \end{aligned}$$

Taking into account that we have uniformly discretized the interval $[0, 1]$ in $n + 1$ subintervals, we get that the value $\left(\frac{2j+1}{2(n+1)}\right)^{-\alpha}\frac{1}{n+1}$ is the (Lebesgue) measure of the rectangle with x-edges $[\frac{j}{n+1}, \frac{j+1}{n+1}]$ and y-edges $\left[0, \left(\frac{2j+1}{2(n+1)}\right)^{-\alpha}\right]$. Therefore, the above sum is approximated by an integral as follows:

$$S_i \approx (n+1)\int_{\frac{i}{n+1}}^{1} x^{-\alpha}dx = \frac{n+1}{1-\alpha}\left[x^{1-\alpha}\right]_{\frac{i}{n+1}}^{1} = \frac{n+1}{1-\alpha}\left[1 - \left(\frac{i}{n+1}\right)^{1-\alpha}\right]. \tag{16}$$

It is easily checked that the error of the above approximation is less than S_i in order of magnitude. If we substitute $i = 0$ in relation (16), then we estimate the quantity S as

$$S \approx (n+1)\int_{0}^{1} x^{-\alpha}dx = \frac{n+1}{1-\alpha}\left[x^{1-\alpha}\right]_{0}^{1} = \frac{n+1}{1-\alpha}. \tag{17}$$

From (16) and (17) we find

$$S - S_i \approx \frac{n+1}{1-\alpha} - \frac{n+1}{1-\alpha}\left[1 - \left(\frac{i}{n+1}\right)^{1-\alpha}\right] = \frac{(n+1)^{\alpha}\ i^{1-\alpha}}{1-\alpha}. \tag{18}$$

By taking the sum of the coefficients in (16), we deduce

$$\begin{aligned} \textstyle\sum_{i=k+1}^{n} S_i &\approx \textstyle\sum_{i=k+1}^{n} \frac{n+1}{1-\alpha} - \sum_{i=k+1}^{n} \frac{n+1}{1-\alpha}\left(\frac{i}{n+1}\right)^{1-\alpha} \\ &\approx \tfrac{(n+1)(n-k)}{1-\alpha} - \tfrac{(n+1)^2}{1-\alpha}\textstyle\int_{\frac{k+1}{n+1}}^{1} x^{1-\alpha}dx \\ &= \tfrac{(n+1)(n-k)}{1-\alpha} - \tfrac{(n+1)^2}{(1-\alpha)(2-\alpha)}\left[1 - \left(\tfrac{k+1}{n+1}\right)^{2-\alpha}\right] \\ &= \tfrac{(n+1)^2}{2-\alpha} + \tfrac{(n+1)^{\alpha}(k+1)^{2-\alpha}}{(1-\alpha)(2-\alpha)} - \tfrac{(n+1)(k+1)}{(1-\alpha)}, \end{aligned} \tag{19}$$

where we used $\sum_{i=k+1}^{n}\left(\frac{i}{n+1}\right)^{1-\alpha}\frac{1}{n+1} \approx \int_{\frac{k+1}{n+1}}^{1} x^{1-\alpha}dx$. Similarly, by taking the sum of the coefficients in (18), we find

$$\begin{aligned} \textstyle\sum_{i=1}^{k}(S-S_i) &\approx \tfrac{(n+1)^{\alpha}}{1-\alpha}\textstyle\sum_{i=1}^{k} i^{1-\alpha} \approx \tfrac{(n+1)^2}{1-\alpha}\int_0^{\frac{k}{n+1}} x^{1-\alpha}dx \\ &= \tfrac{(n+1)^2}{(1-\alpha)(2-\alpha)}\left(\tfrac{k}{n+1}\right)^{2-\alpha} = \tfrac{(n+1)^{\alpha}k^{2-\alpha}}{(1-\alpha)(2-\alpha)}, \end{aligned} \tag{20}$$

where $\sum_{i=1}^{k}\left(\frac{i}{n+1}\right)^{1-\alpha}\frac{1}{n+1} \approx \int_0^{\frac{k}{n+1}} x^{1-\alpha}dx$. By replacing the explicit formulae (16), (17), (18), (19), and (20) in relation (15), we arrive to estimate $\tilde{S}_k$ that is

$$\begin{aligned} \tilde{S}_k =& \left[1-\left(\tfrac{k}{n+1}\right)^{1-\alpha}\right]\tfrac{(n+1)^{\alpha}k^{2-\alpha}}{(1-\alpha)(2-\alpha)} \\ &+ \left(\tfrac{k}{n+1}\right)^{1-\alpha}\left[\tfrac{(n+1)^2}{2-\alpha} + \tfrac{(n+1)^{\alpha}(k+1)^{2-\alpha}}{(1-\alpha)(2-\alpha)} - \tfrac{(n+1)(k+1)}{(1-\alpha)}\right]. \end{aligned} \tag{21}$$

We plainly observe that $\tilde{S}_k$ does not exceed, in order of magnitude, the value $\max\{(n+1)^{\alpha}k^{2-\alpha},\ (n+1)^{1+\alpha}k^{1-\alpha},\ (n+1)^{2\alpha-1}k^{3-2\alpha}\}$. In any case this maximum is of order of n^2. On the other hand, by studying (21) for $\frac{n}{4}+1 \le k \le \frac{3n}{4}$, we obtain that

$$\tilde{S}_k \sim n^2, \qquad \frac{n}{4}+1 \le k \le \frac{3n}{4}. \tag{22}$$

We consider now the matrix $B_{\frac{n}{2}}$, the $\frac{n}{2} \times \frac{n}{2}$ block of A_n^{-1} formed by deleting the first and the last $\frac{n}{4}$ rows and columns. We denote by $\tilde{S}_{B_k}$ the kth row sum of the matrix $B_{\frac{n}{2}}$, where the index k ranges from $\frac{n}{4}+1$ to $\frac{3n}{4}$. Taking into account (15), we infer

$$\tilde{S}_{B_k} = \frac{S_k}{S}\sum_{i=\frac{n}{4}+1}^{k}(S-S_i) + \frac{S-S_k}{S}\sum_{i=k+1}^{\frac{3n}{4}} S_i. \tag{23}$$

By making analogous calculations, as in the estimation of $\tilde{S}_k$, we find

$$\begin{aligned} \tilde{S}_{B_k} =& \left[1-\left(\tfrac{k}{n+1}\right)^{1-\alpha}\right]\tfrac{(n+1)^{\alpha}\left(k^{2-\alpha}-(\frac{n}{4})^{2-\alpha}\right)}{(1-\alpha)(2-\alpha)} \\ &+ \left(\tfrac{k}{n+1}\right)^{1-\alpha}\left[\tfrac{(n+1)(\frac{3n}{4}-k)}{1-\alpha} + \tfrac{(n+1)^{\alpha}\left((k+1)^{2-\alpha}-(\frac{3n}{4}+1)^{2-\alpha}\right)}{(1-\alpha)(2-\alpha)}\right]. \end{aligned} \tag{24}$$

It is easy to understand that (24) implies

$$\tilde{S}_{B_k} \sim n^2, \qquad \frac{n}{4}+1 \le k \le \frac{3n}{4}. \tag{25}$$

We apply now a permutation transformation to the matrix A_n^{-1} in such a way that its block $B_{\frac{n}{2}}$ will appear in the first $\frac{n}{2} \times \frac{n}{2}$ rows and columns. Then, the permuted matrix is dominated to the first $\frac{n}{2} \times \frac{n}{2}$ block, with all the first $\frac{n}{2}$ row sums being of order n^2. In this case Lemma 4.3 is applied to obtain relation (14) that is

$$\kappa(A_n) \sim \rho(A_n^{-1}) \sim n^2.$$

Case 3: $\alpha = 1$.

We follow the same steps as in the previous case:

$$\begin{aligned} S_i =& \sum_{j=i}^{n} \frac{1}{a_{j+\frac{1}{2}}} = \sum_{j=i}^{n} \frac{2(n+1)}{2j+1} \\ &\approx (n+1) \int_{\frac{i}{n+1}}^{1} x^{-1} dx = (n+1) \log\left(\frac{n+1}{i}\right); && (26) \\ S =& 2(n+1) + S_1 \approx (n+1)\left(2 + \log(n+1)\right); && (27) \\ S - S_i \approx& (n+1)\left(2 + \log(n+1)\right) - (n+1) \log\left(\frac{n+1}{i}\right) \\ =& (n+1)\left(2 + \log(i)\right). && (28) \end{aligned}$$

Now by substituting (26), (27), (28) in relation (15), we infer the following estimate for $\tilde{S}_k$:

$$\begin{aligned} \tilde{S}_k \approx& \tfrac{(n+1)\log\left(\frac{n+1}{k}\right)}{(n+1)(2+\log(n+1))} \textstyle\sum_{i=1}^{k} (n+1)\left(2+\log(i)\right) \\ +& \tfrac{(n+1)(2+\log(k))}{(n+1)(2+\log(n+1))} \textstyle\sum_{i=k+1}^{n} (n+1) \log\left(\frac{n+1}{i}\right) \\ =& \tfrac{(n+1)\log\left(\frac{n+1}{k}\right)}{2+\log(n+1)} \left(2k + \textstyle\sum_{i=1}^{k} \log(i)\right) \\ +& \tfrac{(n+1)(2+\log(k))}{2+\log(n+1)} \left((n-k)\log(n+1) - \textstyle\sum_{i=k+1}^{n} \log(i)\right). \end{aligned} \tag{29}$$

On the other hand

$$\sum_{i=1}^{k} \log(i) \approx \int_{1}^{k} \log(x) dx = k\log(k) - k + 1$$

and

$$\sum_{i=k+1}^{n} \log(i) \approx \int_{k+1}^{n} \log(x) dx = n\log(n) - n - k\log(k) + k + 1.$$

By replacing the latter terms in (29), we obtain

$$\begin{aligned} \tilde{S}_k \approx& \tfrac{(n+1)\log\left(\frac{n+1}{k}\right)}{2+\log(n+1)} (k + k\log(k) + 1) \\ +& \tfrac{(n+1)(2+\log(k))}{2+\log(n+1)} \left(n\log\left(\tfrac{n+1}{n}\right) - k\log\left(\tfrac{n+1}{k}\right) + n - k - 1\right), \end{aligned} \tag{30}$$

and hence the quantity $\tilde{S}_k$ does not exceed n^2 in order of magnitude. Furthermore, by analyzing (30) for $\frac{n}{4} + 1 \leq k \leq \frac{3n}{4}$ we obtain relation (22) that is

$$\tilde{S}_k \sim n^2, \qquad \frac{n}{4} + 1 \leq k \leq \frac{3n}{4}.$$

As in the previous case, we consider the matrix $B_{\frac{n}{2}}$, the same $\frac{n}{2} \times \frac{n}{2}$ block of A_n^{-1}, and we estimate the row sums $\tilde{S}_{B_k}$, $\quad \frac{n}{4}+1 \leq k \leq \frac{3n}{4}$, *i.e.*,

$$\begin{aligned} \tilde{S}_{B_k} \approx\ & \tfrac{(n+1)\log\left(\frac{n+1}{k}\right)}{2+\log(n+1)} \left(k - \tfrac{n}{4} + k\log(k) - \tfrac{n}{4}\log\left(\tfrac{n}{4}\right)\right) \\ & + \tfrac{(n+1)(2+\log(k))}{2+\log(n+1)} \left(\tfrac{3n}{4} - k + \tfrac{3n}{4}\log\left(\tfrac{n+1}{\frac{3n}{4}}\right) + k\log(\tfrac{n+1}{k})\right). \end{aligned} \tag{31}$$

It is easily checked that (31) implies the same conclusion (25), as in the previous case. Applying again Lemma 4.3 as in (14), we find

$$\kappa(A_n) \sim \rho(A_n^{-1}) \sim n^2.$$

Case 4: $1 < \alpha < 2$.

In analogy with the previous cases we estimate

$$S_i = \sum_{j=i}^{n} \left(\frac{2j+1}{2(n+1)}\right)^{-\alpha} \approx (n+1)\int_{\frac{i}{n+1}}^{1} x^{-\alpha}dx = \frac{n+1}{\alpha-1}\left[\left(\frac{n+1}{i}\right)^{\alpha-1} - 1\right]; \tag{32}$$

$$S = 2^\alpha(n+1)^\alpha + S_1 \approx (n+1)^\alpha\left[2^\alpha + \frac{1}{\alpha-1}\right] - \frac{n+1}{\alpha-1}; \tag{33}$$

$$S - S_i \approx (n+1)^\alpha\left[2^\alpha + \frac{1}{\alpha-1}\left(1 - \frac{1}{i^{\alpha-1}}\right)\right]; \tag{34}$$

$$\begin{aligned} \textstyle\sum_{i=k+1}^{n} S_i &\approx \tfrac{(n+1)^\alpha}{\alpha-1}\textstyle\sum_{i=k+1}^{n} i^{1-\alpha} - \tfrac{(n+1)(n-k)}{\alpha-1} \\ &\approx \tfrac{(n+1)^2}{\alpha-1}\textstyle\int_{\frac{k+1}{n+1}}^{1} x^{1-\alpha}dx - \tfrac{(n+1)(n-k)}{\alpha-1} \\ &= \tfrac{1}{\alpha-1}\left[\tfrac{1}{2-\alpha}(n+1)^2 + (n+1)(k+1) - \tfrac{(n+1)^\alpha(k+1)^{2-\alpha}}{(2-\alpha)}\right]; \end{aligned} \tag{35}$$

$$\begin{aligned} \textstyle\sum_{i=1}^{k}(S - S_i) &\approx (n+1)^\alpha k\left(2^\alpha + \tfrac{1}{\alpha-1}\right) - \tfrac{(n+1)^\alpha}{\alpha-1}\textstyle\sum_{i=1}^{k} i^{1-\alpha} \\ &\approx (n+1)^\alpha k\left(2^\alpha + \tfrac{1}{\alpha-1}\right) - \tfrac{(n+1)^2}{\alpha-1}\textstyle\int_0^{\frac{k}{n+1}} x^{1-\alpha}dx \\ &= (n+1)^\alpha k\left(2^\alpha + \tfrac{1}{\alpha-1} - \tfrac{k^{1-\alpha}}{(\alpha-1)(2-\alpha)}\right). \end{aligned} \tag{36}$$

Substituting the explicit quantities (32), (33), (34), (35), and (36) in relation (15), we deduce that

$$\begin{aligned} \tilde{S}_k \approx\ & \tfrac{\frac{n+1}{\alpha-1}\left[\left(\frac{n+1}{k}\right)^{\alpha-1}-1\right]}{(n+1)^\alpha\left[2^\alpha+\frac{1}{\alpha-1}\right]-\frac{n+1}{\alpha-1}}(n+1)^\alpha k\left(2^\alpha + \tfrac{1}{\alpha-1} - \tfrac{k^{1-\alpha}}{(\alpha-1)(2-\alpha)}\right) \\ & + \tfrac{(n+1)^\alpha\left[2^\alpha+\frac{1}{\alpha-1}\left(1-\frac{1}{k^{\alpha-1}}\right)\right]}{(n+1)^\alpha\left[2^\alpha+\frac{1}{\alpha-1}\right]-\frac{n+1}{\alpha-1}}\left[\tfrac{(n+1)^2}{(\alpha-1)(2-\alpha)} + \tfrac{(n+1)(k+1)}{\alpha-1} - \tfrac{(n+1)^\alpha(k+1)^{2-\alpha}}{(\alpha-1)(2-\alpha)}\right]. \end{aligned} \tag{37}$$

A plain analysis of the main terms of (37) shows that the order of $\tilde{S}_k$ does not exceed n^2. Moreover, the study of (30) for $\frac{n}{4}+1 \leq k \leq \frac{3n}{4}$ leads to relation (22), i.e.,

$$\tilde{S}_k \sim n^2, \qquad \frac{n}{4}+1 \leq k \leq \frac{3n}{4}.$$

We consider once again the matrix $B_{\frac{n}{2}}$. Then

$$\begin{aligned}\tilde{S}_{B_k} \approx\ & \frac{\frac{n+1}{\alpha-1}\left[\left(\frac{n+1}{k}\right)^{\alpha-1}-1\right](n+1)^{\alpha}}{(n+1)^{\alpha}\left[2^{\alpha}+\frac{1}{\alpha-1}\right]-\frac{n+1}{\alpha-1}}\left[\left(k-\frac{n}{4}\right)\left(2^{\alpha}+\frac{1}{\alpha-1}\right)-\frac{k^{2-\alpha}-\left(\frac{n}{4}\right)^{2-\alpha}}{(\alpha-1)(2-\alpha)}\right] \\ & + \frac{(n+1)^{\alpha}\left[2^{\alpha}+\frac{1}{\alpha-1}\left(1-\frac{1}{k^{\alpha-1}}\right)\right]}{(n+1)^{\alpha}\left[2^{\alpha}+\frac{1}{\alpha-1}\right]-\frac{n+1}{\alpha-1}}\left[\frac{(n+1)^{\alpha}\left(\left(\frac{3n}{4}+1\right)^{2-\alpha}-k^{2-\alpha}\right)}{(\alpha-1)(2-\alpha)}-\frac{(n+1)\left(\frac{3n}{4}-k\right)}{\alpha-1}\right].\end{aligned} \tag{38}$$

The analysis of (38) gives the same conclusion as (25), and then, by Lemma 4.3, we obtain relation (14), i.e., $\kappa(A_n) \sim \rho(A_n^{-1}) \sim n^2$.

Case 5: $\alpha = 2$.

As in the preceding cases we have:

$$S_i = \sum_{j=i}^{n}\left(\frac{2j+1}{2(n+1)}\right)^{-2} \approx (n+1)\int_{\frac{i}{n+1}}^{1} x^{-2}dx = (n+1)\left(\frac{n+1}{i}-1\right); \tag{39}$$

$$S = 4(n+1)^2 + S_1 \approx (n+1)(5n+4); \tag{40}$$

$$S - S_i \approx (n+1)(5n+4) - (n+1)\left(\frac{n+1}{i}-1\right) = (n+1)^2\left(5-\frac{1}{i}\right); \tag{41}$$

$$\begin{aligned}\textstyle\sum_{i=k+1}^{n} S_i &\approx (n+1)\textstyle\sum_{i=k+1}^{n}\left(\frac{i}{n+1}\right)^{-1} - (n+1)(n-k) \\ &\approx (n+1)^2 \textstyle\int_{\frac{k+1}{n+1}}^{1} x^{-1}dx - (n+1)(n-k) \\ &= (n+1)^2 \log\left(\frac{n+1}{k+1}\right) - (n+1)(n-k);\end{aligned} \tag{42}$$

$$\begin{aligned}\textstyle\sum_{i=1}^{k}(S-S_i) &\approx 5(n+1)^2k - (n+1)\textstyle\sum_{i=1}^{k}\left(\frac{i}{n+1}\right)^{-1} \\ &\approx 5(n+1)^2k - (n+1)^2 \textstyle\int_{\frac{1}{n+1}}^{\frac{k+1}{n+1}} x^{-1}dx \\ &= (n+1)^2\left(5k - \log(k+1)\right).\end{aligned} \tag{43}$$

For the estimation of $\tilde{S}_k$ we employ (39), (40), (41), (42), and (43) in relation (15):

$$\begin{aligned}\tilde{S}_k &\approx \frac{\frac{n+1}{k}-1}{5n+4}(n+1)^2\left(5k-\log(k+1)\right) \\ &+ \frac{(n+1)\left(5-\frac{1}{k}\right)}{5n+4}(n+1)\left((n+1)\log\left(\frac{n+1}{k+1}\right)-(n-k)\right) \\ &= \frac{(n+1)^2}{5n+4}[5 + \frac{n-k}{k} + 5(n+1)\log(n+1) \\ &- (5n+4)\log(k+1) - \frac{n+1}{k}\log(n+1)].\end{aligned} \tag{44}$$

A straightforward conclusion is that $\tilde{S}_k$ does not exceed $n^2\log(n)$ in order of magnitude. On the other hand, by exploiting (44) for $1 \leq k \leq m$, where m is a constant integer independent of n, we obtain

$$\tilde{S}_k \sim n^2\log(n), \qquad 1 \leq k \leq m. \tag{45}$$

By the Perron Frobenius theory on nonnegative matrices, we find

$$\rho(A_n^{-1}) = O(n^2\log(n)). \tag{46}$$

We consider the $m \times m$ matrix B_m, which is the submatrix of A_n^{-1} formed by the first m rows and columns. The estimation of the row sums $\tilde{S}_{B_k}, \quad 1 \leq k \leq m$, leads to

$$\begin{aligned} \tilde{S}_{B_k} &\approx \tfrac{\frac{n+1}{k}-1}{5n+4}(n+1)^2\left(5k-\log(k+1)\right) \\ &+ \tfrac{(n+1)\left(5-\frac{1}{k}\right)}{5n+4}(n+1)\left((n+1)\log\left(\tfrac{m+1}{k+1}\right)-(m-k)\right) \\ &= \tfrac{(n+1)^2}{5n+4}[5(n+1-m)+\tfrac{m-k}{k}+5(n+1)\log(m+1) \\ &- (5n+4)\log(k+1)-\tfrac{n+1}{k}\log(m+1)]. \end{aligned} \tag{47}$$

Since m and k are constant independent of n, it follows that

$$\tilde{S}_{B_k} \sim n^2, \qquad\qquad 1 \leq k \leq m. \tag{48}$$

By the interlacing law we obtain $\rho(A_n^{-1}) \geq \rho(B_m) \sim n^2$ and therefore

$$\rho(A_n^{-1}) = \Omega(n^2). \tag{49}$$

In conclusion, from (49) and (46), we deduce that

$$\kappa(A_n) \sim \rho(A_n^{-1}) = O(n^2\log(n)) \cap \Omega(n^2).$$

Case 6: $\alpha > 2$.

It is easily seen that the estimation of the quantities $S_i, \quad S, \quad S-S_i$ and $\sum_{i=k+1}^{n} S_i$ is just the same as in Case 4, when dealing with relations (32), (33), (34), and (35), respectively. The only modification we need is to estimate the quantity $\sum_{i=1}^{k} S - S_i$, by exploiting an alternative approximation since $\int_0^{\frac{k}{n+1}} x^{1-\alpha}dx$ diverges for $\alpha > 2$. More in detail we have

$$\begin{aligned} \textstyle\sum_{i=1}^{k}(S-S_i) &\approx (n+1)^\alpha k\left(2^\alpha+\tfrac{1}{\alpha-1}\right)-\tfrac{(n+1)^\alpha}{\alpha-1}\textstyle\sum_{i=1}^{k} i^{1-\alpha} \\ &\approx (n+1)^\alpha k\left(2^\alpha+\tfrac{1}{\alpha-1}\right)-\tfrac{(n+1)^\alpha}{\alpha-1}\int_{\frac{1}{n+1}}^{\frac{k+1}{n+1}} x^{1-\alpha}dx \\ &= (n+1)^\alpha\left[k\left(2^\alpha+\tfrac{1}{\alpha-1}\right)-\tfrac{1-\frac{1}{(k+1)^{\alpha-2}}}{(\alpha-1)(\alpha-2)}\right]. \end{aligned} \tag{50}$$

We estimate $\tilde{S}_k$ by replacing (32), (33), (34), (35), and (50) in relation (15):

$$\begin{aligned} \tilde{S}_k \approx\; & \frac{\frac{n+1}{\alpha-1}\left[\left(\frac{n+1}{k}\right)^{\alpha-1}-1\right]}{(n+1)^\alpha\left[2^\alpha+\frac{1}{\alpha-1}\right]-\frac{n+1}{\alpha-1}}(n+1)^\alpha\left[k\left(2^\alpha+\tfrac{1}{\alpha-1}\right)-\tfrac{1-\frac{1}{(k+1)^{\alpha-2}}}{(\alpha-1)(\alpha-2)}\right] \\ +\; & \frac{(n+1)^\alpha\left[2^\alpha+\frac{1}{\alpha-1}\left(1-\frac{1}{k^{\alpha-1}}\right)\right]}{(n+1)^\alpha\left[2^\alpha+\frac{1}{\alpha-1}\right]-\frac{n+1}{\alpha-1}}\left[\tfrac{(n+1)^\alpha(k+1)^{2-\alpha}}{(\alpha-1)(\alpha-2)}-\tfrac{(n+1)^2}{(\alpha-1)(\alpha-2)}+\tfrac{(n+1)(k+1)}{\alpha-1}\right]. \end{aligned} \tag{51}$$

Again we deduce that $\tilde{S}_k$ grows in order as n^α. Moreover, by studying (51) for $1 \leq k \leq \bar{k}$, where $\bar{k}$ is a constant independent of n, we find that both terms of (51) are of order n^α. Thus

$$\tilde{S}_k \sim n^\alpha, \qquad 1 \leq k \leq \bar{k}. \tag{52}$$

By considering the matrix $B_{\bar{k}}$, the submatrix of A_n^{-1} formed by the first $\bar{k}$ rows and columns, in the formula of the kth row sum of $B_{\bar{k}}$ the first term of (51) appears unchanged, while the changes appear only in the second term. Thus,

$$\tilde{S}_{B_k} \sim n^\alpha, \qquad 1 \leq k \leq \bar{k}. \tag{53}$$

Finally, by Lemma 4.3 we obtain

$$\kappa(A_n) \sim \rho(A_n^{-1}) \sim n^\alpha, \tag{54}$$

and the proof of the theorem is completed. □

5 The case of higher order BVPs

The results of Theorem 4.4 can be extended in a straightforward manner to cover the case where the BVP is of order higher than 2, i.e., our equations are of the form

$$\begin{cases} (-1)^k \frac{d^k}{dx^k}\left(a(x)\frac{d^k}{dx^k}u(x)\right) = f\,(x) \quad \text{on } \Omega = (0,1), \quad k = 2, 3, \ldots \\ \text{homogeneous B.C. on } \ \partial\Omega, \end{cases} \tag{55}$$

where the function $a(x)$ has a root at $\tilde{x}_0 \in \overline{\Omega}$ of order α. In analogy to the case of second order operators, we approximate (55) on a uniform grid of stepsize $h = (n+1)^{-1}$, using centered finite differences of minimal precision order 2. As a consequence we find $2k+1$ band $n \times n$ linear systems $A_n(a)x = b$.

The generalization of Theorem 4.4 takes the following form:
Theorem 5.1. *Let $\{A_n\}_n$, $A_n \in \mathbb{R}^{n \times n}$, be the sequence of matrices derived from the descritization of the Semielliptic Differential Equation (55) with the bounded coefficient function $a(x)$ having a unique root at 0 of order α i.e. $a(x) \sim x^\alpha$ on $D = [0,1]$. Then, for the spectral condition number $\kappa(A_n)$ of the matrix A_n which coincides in order with the spectral radius of A_n^{-1}, there holds:*

$$\kappa_2(A_n) \sim \begin{cases} n^{2k}, & 0 \leq \alpha < 2k, \\ O(n^{2k}\log(n)) \cap \Omega(n^{2k}), & \alpha = 2k, \\ n^\alpha, & \alpha > 2k. \end{cases} \tag{56}$$

The proof follows exactly the same governing ideas as the proof of Theorem 4.4 but with the mathematical manipulation becoming more and more complicated and tricky as the order of the BVP increases. The reason for that concerns essentially the formulation of the explicit form for the inverse of the coefficient matrix A_n. In Section 7 we give many numerical examples regarding the case of BVPs with order higher than two, with all of them fully confirming the theoretical results given in Theorem 5.1.

Remark 5.1. *The assumption for the coefficient function regarding the uniqueness of its root cannot be relaxed to "many isolated roots". The reason, which has been mentioned also in* [25], *is that in this case the condition number grows in an unpredictable (nonmonotone) way as the dimension of the problem tends to infinity: in reality, the matrix A_n may happen to be also singular for certain dimensions. More specifically, performing various numerical experiments (see Tables 9, 10, and 11 for a partial account on our findings), we have observed that:*

- *For k=1,2,3 there exists $a(x)$ such that $\max\{\alpha_i\} < 2k$ and $\kappa_2(A_n) \not\sim n^{2k}$; more precisely, $\kappa_2(A_n) = \Omega(n^{2k+\delta})$, for some $\delta > 0$;*
- *For k=1,2,3 there exists $a(x)$ such that $\max\{\alpha_i\} = 2k$ and $\kappa_2(A_n) \not\sim O(n^{2k}\log(n)) \cap \Omega(n^{2k})$; more in detail, $\kappa_2(A_n) = \Omega(n^{2k+\delta})$, for some $\delta > 0$;*
- *For k=1,2,3 there exists $a(x)$ such that $\max\{\alpha_i\} > 2k$ and $\kappa_2(A_n) \not\sim n^{\max\{\alpha_i\}}$; more precisely, $\kappa_2(A_n) = \Omega(n^{\max\{\alpha_i\}+\delta})$, for some $\delta > 0$.*

In Section 7 we report some examples concerning this case, and the conclusion is that the condition numbers grow faster, when compared with the bounds in Theorem 4.4 and Theorem 5.1: the reason is a kind of interference between the sources of ill-conditioning represented by the different zeros (for a nice contrast with the case of a unique zero, see the discussion at the end of Subsection 2.2).

6 Remarks on the 2D case

We consider the 2D problem

$$-\frac{\partial}{\partial x}\left(a(x,y)\frac{\partial}{\partial x}u\right) - \frac{\partial}{\partial y}\left(b(x,y)\frac{\partial}{\partial y}u\right) = f(x,y) \qquad (57)$$

with Dirichlet boundary conditions. Using the well-known five points formula and by ordering the unknowns in the classic manner, we arrive to the $n^2 \times n^2$ linear system

$$A_{nn}x = b,$$

where A_{nn} is a symmetric positive definite block tridiagonal matrix, with the diagonal blocks being tridiagonal matrices and the off diagonal blocks being diagonal ones.

As we have mentioned from the beginning of this paper, the main contribution of this work will be to give a guideline and to establish a theoretical framework for dealing with the more interesting 2D case, which is of great importance from both, theoretical and practical point of view. A trivial but immediate application of our estimation of the condition number to the 2D case, is the circumstance where the coefficient functions are of separable variables. In addition, we perform various numerical experiments and it clearly emerges that the results of Theorem 4.4, under suitable assumptions, can be analogously extended to cover also the 2D case. The following definition is useful.

Definition 6.1. *Let $f(x,y)$ be a nonnegative bounded function having a zero at (x_0,y_0). We say that the order of zero is $\alpha \in (0,\infty)$ if there exists a finite number p of curves C_i, $i=1,\dots,p$, defined by $l_i(x,y)=0$, passing through (x_0,y_0) and regular in it such that $f \sim \hat{f}$ and*

$$\hat{f}(x,y)=\sum_{i=1}^{p}|l_i(x,y)|^{\alpha}+g(x,y),$$

where g has a zero at (x_0,y_0) of order at least $\beta>\alpha$.

We are ready to state our conjecture concerning the relation of the condition of A_{nn} and the order of the zeros of the coefficient functions:

Statement 6.1. *Let us assume that the coefficient functions $a(x,y), b(x,y)$ have zeros $(x_0,y_0),(x_1,y_1)$ of orders α_a,α_b, respectively. Then for the spectral condition number $\kappa_2(A_{nn})$ of the matrix A_{nn} there holds:*

$$\kappa_2(A_{nn}) \sim \begin{cases} n^2, & 0 \le \min\{\alpha_a,\alpha_b\} < 2; \\ O(n^2\log(n)) \cap \Omega(n^2), & \min\{\alpha_a,\alpha_b\}=2, \\ n^{\min\{\alpha_a,\alpha_b\}}, & \min\{\alpha_a,\alpha_b\}>2. \end{cases}$$

7 Numerical experiments

In this section we present several numerical tests concerning both 1D and 2D BVPs. We will start by discussing experiments on univariate BVPs of order 2, 4, and 6, respectively.

The quantity which is of main interest in our context is the estimation of

$$\rho_m=\log_2\left(\frac{\lambda_{\min}(A_{2^m})}{\lambda_{\min}(A_{2^{(m+1)}})}\right).$$

We observe that ρ_m reflects the decrement rate of the minimal eigenvalue of the coefficient matrix A_n.

For the second order BVP in (1) we have used as coefficient functions the following test functions:

$$a_1(x) = \left|x - \frac{1}{\sqrt{2}}\right|, \quad a_2(x) = (x - .3)^2, \quad a_3(x) = \left|x - \frac{\pi}{4}\right|^{\frac{5}{2}}$$

and the results are given in Tables 1, 2, and 3, respectively. Regarding the fourth order BVP i.e. (55) with $k = 2$, we use the functions

$$a_4(x) = \left(x - \frac{1}{\sqrt{3}}\right)^2, \quad a_5(x) = \sin(x)^4, \quad a_6(x) = x^5,$$

with associated results in Tables 4, 5, and 6, while, for the sixth order BVP i.e. (55) with $k = 3$, we have chosen as coefficient functions

$$a_7(x) = \sin(x)^4, \quad a_8(x) = x^7,$$

with related results in Tables 7, and 8.

Obviously, in order to perform a meaningful test for our theoretical derivations, the considered coefficient functions have different analytical behaviors, and with roots of order less, equal or greater than the order of the differential equation. In all cases, we ascertain numerically the theoretical findings in Theorems 4.4 and 5.1.

Table 1
1D, $k = 1$:$a(x) = \left|x - \frac{1}{\sqrt{2}}\right|$.

m	5	6	7	8	9	10	11	12	13	14
$\lambda_{\min}$	2.263×10^{-3}	5.483×10^{-4}	1.324×10^{-4}	3.2×10^{-5}	7.76×10^{-6}	1.889×10^{-6}	4.612×10^{-7}	1.129×10^{-7}	2.773×10^{-8}	6.824×10^{-9}
ρ_m	2.045	2.05	2.048	2.044	2.04	2.034	2.03	2.026	2.023	

For the case where $a(x)$ has multiple roots in $[0, 1]$ things completely change as reported in Remark 5.1. Tables 9, 10, and 11 show this "irregular" behavior for the quantity ρ_m.

a) $k = 2 \quad a(x) = x^3|x - .3|^{\frac{5}{2}}$,
b) $k = 2 \quad a(x) = (x - \frac{1}{\sqrt{2}})^2(x - \frac{1}{\sqrt{3}})^4$,
c) $k = 1 \quad a(x) = (x - .5)^2 x^3$.

For the 2D case, we consider the following four examples:

a) $a(x, y) = b(x, y) = x + y$,

Table 2
1D, $k = 1$:$a(x) = (x - .3)^2$.

m	5	6	7	8	9
$\lambda_{\min}$	3.251×10^{-4}	9.608×10^{-5}	2.063×10^{-5}	4.882×10^{-6}	1.179×10^{-6}
ρ_m	1.759	2.22	2.079	2.050	1.898
m	10	11	12	13	14
$\lambda_{\min}$	3.163×10^{-7}	7.282×10^{-8}	1.762×10^{-8}	4.318×10^{-9}	1.123×10^{-9}
ρ_m	2.119	2.047	2.029	1.943	

Table 3
1D, $k = 1$:$a(x) = (x - \pi/4)^{\frac{5}{2}}$.

m	5	6	7	8	9
$\lambda_{\min}$	5.206×10^{-5}	9.278×10^{-6}	2.46×10^{-6}	3.296×10^{-7}	5.446×10^{-8}
ρ_m	2.488	1.915	2.9	2.597	2.549
m	10	11	12	13	14
$\lambda_{\min}$	9.305×10^{-9}	2.282×10^{-9}	3.633×10^{-10}	6.529×10^{-11}	1.193×10^{-11}
ρ_m	2.028	2.651	2.476	2.452	

Table 4
1D, $k = 2$:$a(x) = \left|x - \frac{1}{\sqrt{3}}\right|$.

m	5	6	7	8	9	10	11	12	13
ρ_m	3.646	3.961	3.996	4.038	3.995	4.151	3.851	4.022	4.034

Table 5
1D, $k = 2$: $a(x) = \sin(x)^4$.

m	5	6	7	8	9	10	11	12	13
ρ_m	4.208	4.189	4.161	4.135	4.113	4.094	4.078	4.066	4.056

Table 6
1D, $k = 2$: $a(x) = x^5$.

m	5	6	7	8	9	10	11	12	13
ρ_m	4.911	4.951	4.974	4.987	4.993	4.997	4.998	4.999	5

b) $a(x, y) = x^3 + y^4$, $b(x, y) = x^5 + y^6$,
c) $a(x, y) = x^2 + y^2$, $b(x, y) = (x + y)^2$,
d) $a(x, y) = |x - y|^3$, $b(x, y) = |x - \frac{1}{2}|^3 + |y - \frac{1}{2}|^3$.

The results in Tables 12, 13, 14, and 15 fully confirm the statements formulated

Table 7

1D, $k=3$: $a(x)=\sin(x)^4$.

m	5	6	7	8	9	10	11	12	13
ρ_m	5.853	5.929	5.965	5.983	5.992	5.996	5.999	5.999	6

Table 8

1D, $k=3$: $a(x)=x^7$.

m	5	6	7	8	9	10	11	12	13
ρ_m	6.846	6.922	6.961	6.980	6.990	6.995	6.998	6.999	6.999

Table 9

1D, $k=2$, multiple root case: $a(x)=x^3|x-.3|^{\frac{5}{2}}$.

m	5	6	7	8	9	10	11	12	13
ρ_m	5.51	2.982	2.278	3.574	7.025	1.811	1.779	3.52	6.947

Table 10

1D, $k=2$, multiple root case : $a(x)=\left(x-\frac{1}{\sqrt{2}}\right)^2\left(x-\frac{1}{\sqrt{3}}\right)^4$.

m	5	6	7	8	9	10	11	12	13
ρ_m	7.264	4.516	4.601	4.316	5.192	4.719	6.447	3.725	4.876

Table 11

1D, $k=2$ multiple root case: $a(x)=(x-.5)^2x^3$.

m	5	6	7	8	9	10	11	12	13
ρ_m	3.89	3.944	3.972	3.986	3.993	3.997	3.998	3.999	4

at the end of Section 6.

Table 12

2D case: $a(x,y)=b(x,y)=x+y$

m	3	4	5	6	7	8
ρ_m	1.826	1.911	1.956	1.978	1.989	1.995

Table 13

2D case: $a(x,y)=x^2+y^2$, $b(x,y)=(x+y)^2$

m	3	4	5	6	7	8
ρ_m	1.921	1.967	1.990	2.001	2.005	2.008

Table 14
2D case: $a(x,y) = x^3 + y^4,\ b(x,y) = x^5 + y^6$

m	3	4	5	6	7	8
ρ_m	2.885	2.949	2.979	2.992	2.997	2.999

Table 15
2D Case: $a(x,y) = |x-y|^3,\ b(x,y) = |x-\frac{1}{2}|^3 + |y-\frac{1}{2}|^3$

m	3	4	5	6	7	8
ρ_m	2.757	2.871	2.934	2.967	2.983	2.991

8 Conclusions

In this paper we have studied the conditioning of semi-elliptic differential problems (the elliptic case is plain thanks to monotonicity arguments). As a main tool we have employed the notion of positivity in three different aspects: definite positivity, operator positivity (especially in Subsection 2.2), and component-wise positivity (especially in Section 4). Our main result is that the two sources of ill-conditioning, the low frequencies coming from the constant coefficient Laplacian, and the space spanned by few canonical vectors related to the position of the zero of $a(x)$, do not interfere; conversely, we numerically observe a bad interference, a kind of resonance, in presence of distinct zeros in the coefficient $a(x)$. Therefore, when a unique zero is considered, there is only a superposition effect so that the size of the degenerating subspace, i.e. that related to small eigenvalues, becomes larger, but the order of ill-conditioning is not worse than that of the two factors separately. As a consequence, both for designing multigrid methods or preconditioners, we can treat the two ill-conditioned spaces separately and this of course implies a simplification in the practical programming and in the theoretical convergence analysis (see e.g. [21,24,26,4])). Finally, there is still the open problem of completing our study in three directions: we would like to identify the constants hidden in the equivalence relations of the main Theorems 4.4 and 5.1, we would like to add more terms if the asymptotic expansion of the condition number of A_n, and, more important, we would like to include the more challenging multidimensional setting. Indeed, as a final remark, we stress that partial results are easily available, by repeating e.g. the same derivations as in Subsection 2.2 in a multilevel setting: however a complete analysis is still missing.

References

[1] D. Aregba-Driollet, R. Natalini, and S. Tang, *Explicit diffusive kinetic schemes for nonlinear degenerate parabolic systems*, Math. Comp. **73-245** (2004), 63–94.

[2] W. Barrett, *A theorem on inverse of tridiagonal matrices*, Linear Algebra Appl. **27** (1979), 211–217.

[3] B. Beckermann and S. Serra Capizzano, *On the asymptotic spectrum of Finite Elements matrices*, SIAM J. Numer. Anal. (to appear).

[4] D. Bertaccini, G. Golub, S. Serra Capizzano, and C. Tablino Possio, *Preconditioned HSS methods for the solution of non-Hermitian positive definite linear systems and applications to the discrete convection-diffusion equation*, Numer. Math. **99** (2005), 441–484.

[5] R. Bevilacqua and M. Capovani, *Proprietà delle matrici a banda ad elementi ed a blocchi*, Bolletino UMI **13-B** (1976), 844–861.

[6] R. Bhatia, *Matrix Analysis*, Springer-Verlag, New York, 1997.

[7] A. Böttcher and S. Grudsky, *Spectral Properties of Banded Toeplitz Matrices*, SIAM Publ., Philadelphia, PA, 2005.

[8] M. Capovani, *Sulla determinazione della inversa delle matrici tridiagonali e tridiagonali a blocchi*, Calcolo **7** (1970), 295–303.

[9] ———, *Su alcune propiertà delle matrici tridiagonali e pentadiagonali*, Calcolo **8** (1971), 149–159.

[10] R. Chan and T. Chan, *Circulant preconditioners for elliptic problems*, J. Numer. Linear Algebra Appl. **1** (1992), 77–101.

[11] G. Fiorentino and S. Serra Capizzano, *Fast parallel solvers for elliptic problems*, Comput. Math. Appl. **32** (1996), 61–68.

[12] ———, *Multigrid methods for symmetric positive definite block Toeplitz matrices with nonnegative generating functions*, SIAM J. Sci. Comput. **17** (1996), no. 4, 1068–1081.

[13] F. Gantmacher and M. Krein, *Oszillationsmatrizen, oszillationskerne und kleine schwingungen mechanischer systeme*, Akademie-Verlag, Berlin, 1960.

[14] G. Golub and C. Van Loan, *Matrix Computations*, Johns Hopkins University Press, Baltimore, London, 1996.

[15] A. Greenbaum, *Iterative Methods for Solving Linear Systems*, SIAM Publ., Philadelphia, PA, 1997.

[16] D. Marini and P. Pietra, *Mixed finite element approximation of a degenerate elliptic problem*, Numer. Math. **71** (1995), 225–236.

[17] J. McDonald, R. Nabben, M. Neumann, Schneider H., and M. Tsatsomeros, *Inverse tridiagonal Z-matrices*, Linear and Multilinear Algebra **45** (1998), 75–97.

[18] R. Nabben, *Decay rates of the inverse of nonsymmetric tridiagonal and band matrices*, SIAM J. Matrix Anal. Appl. **20** (1999), 820–837.

[19] P. Rózsa, *On the inverse of band matrices*, Integral equations Operator Theory **10** (1987), 82–95.

[20] S. Serra Capizzano, *Multi-iterative methods*, Comput. Math. Appl. **26** (1993), 65–87.

[21] ———, *The rate of convergence of Toeplitz based PCG methods for second order nonlinear boundary value problems*, Numer. Math. **81** (1999), 461–495.

[22] ———, *Locally X matrices, spectral distributions, preconditioning, and applications*, SIAM J. Matrix Anal. Appl. **21** (2000), 1354–1388.

[23] ———, *Some theorems on linear positive operators and functionals and their applications*, Comput. Math. Appl. **39** (2000), 139–167.

[24] ———, *Convergence analysis of two-grid methods for elliptic Toeplitz and PDEs matrix-sequences*, Numer. Math. **92** (2002), 433–465.

[25] S. Serra Capizzano and C. Tablino Possio, *Spectral and structural analysis of high precision finite difference matrices for elliptic operators*, Linear Algebra Appl. **293** (1999), 85–131.

[26] ———, *Multigrid methods for multilevel circulant matrices*, SIAM J. Sci. Comput. **26** (2004), 55–85.

[27] S. Serra Capizzano and P. Tilli, *Extreme singular values and eigenvalues of non Hermitian Toeplitz matrices*, J. Computat. Appl. Math. **108** (1999), no. 1-2, 113–130.

[28] ———, *On unitarily invariant norms of matrix valued linear positive operators*, J. Ineq. Appl. **7** (2002), no. 3, 309–330.

[29] P. Tilli, *Locally Toeplitz sequences: spectral properties and applications*, Linear Algebra Appl. **278** (1998), 91–120.

[30] U. Trottenberg, C.W. Oosterlee, and A. Schüller, *Multigrid*, Academic Press, London, 2001.

[31] R. Varga, *Matrix Iterative Analysis*, Prentice Hall, Englewood Cliffs, NJ, 1962.

[32] P. Wilmott, S. Howison, and J. Dewynne, *The Mathematics of Financial Derivatives*, Cambridge Univ. Press, Cambridge, MA, 1998.