Discrete Maximum Principle and Positivity Certificates for the Bernstein Dual Petrov–Galerkin Method

Tareq Hamadneh, Jochen Merker, and Gregor Schuldt

Abstract In this article, we discuss the validity of the discrete maximum principle for the spectral method called Bernstein-Dual-Petrov-Galerkin method [4] in case of a uniformly elliptic second-order linear partial differential equation (PDE) in divergence form and corresponding Dirichlet boundary values problems on simply connected domains, which have no holes and are therefore diffeomorphic to a cube.

Keywords Discrete maximum principle · Positivity certificates · Bernstein dual Petrov–Galerkin method · Numerical analysis

1 Introduction

Consider Poisson's equation for homogeneous Dirichlet boundary conditions

$$-\Delta u = f \ \text{ in } \Omega\,, \quad u = 0 \ \text{ on } \partial\Omega \tag{1}$$

on a bounded domain $\Omega \subset \mathbb{R}^N$ with boundary $\partial\Omega$ piecewise sufficiently smooth. By the weak maximum principle, $f \geq 0$ implies $u \geq 0$ in Ω for the solution u of (1). [1] proved an analogous discrete maximum principle (DMP) for a finite difference (FD) discretization of (1), and [2] presented a DMP suitable for both finite element (FE) and FD discretizations by providing a practically convenient set of sufficient conditions on matrix blocks implying validity of a DMP. While these conditions imply for a piecewise linear triangular FE discretization of (1) that the inverse of the stiffness matrix is positive under the interior edge condition (the sum of the two

T. Hamadneh
Al Zaytoonah University of Jordan, Airport Rd., Amman, Jordan

J. Merker (✉) · G. Schuldt
Leipzig University of Applied Sciences, PF 30 11 66, 04251 Leipzig, Germany
e-mail: jochen.merker@htwk-leipzig.de

G. Schuldt
e-mail: gregor.schuldt@htwk-leipzig.de

D. Zeidan et al. (eds.), *Mathematics and Computation*, Springer Proceedings in Mathematics & Statistics 418, https://doi.org/10.1007/978-981-99-0447-1_16

angles opposite to every interior edge is $\leq \pi$), the DMP may fail for certain meshes [3].

In this article, we discuss the validity of the discrete maximum principle for the spectral method called Bernstein-Dual-Petrov-Galerkin method [4] in case of a uniformly elliptic second-order linear partial differential equation (PDE) in divergence form and corresponding Dirichlet boundary values problems on simply connected domains $\Omega \subset \mathbb{R}^N$, which have no holes and are therefore diffeomorphic to a cube. This numerical method combines two advantages, the exponential fast convergence of a spectral method in the interior of Ω for analytic data, and the good approximation properties of Bernstein polynomials [5]. Particularly, the latter will allow us to certify the positivity of numerical solutions. For Helmholtz equation subject to homogeneous Neumann boundary conditions and Bernstein Bubnov-Galerkin method, see [6].

1.1 Outline

In Sect. 2, we provide basic information about linear elliptic PDEs in divergence form; about positivity, the maximum principle and the comparison principle for classical and weak solutions; about Bernstein polynomials and the induced dual polynomials resp. the modal basis functions; and about different certificates of non-negativity resp. positivity. In Sect. 3, the Bernstein dual Petrov–Galerkin method is formulated for general linear elliptic PDEs in divergence form on a domain diffeomorphic to a cube. In the main Sect. 4 of this paper, we discuss algebraic and functional discrete maximum principles for this method as well as Bernstein certificates of non-negativity resp. positivity for the approximate solution in a way, which easily generalizes to dual Petrov–Galerkin methods with arbitrary non-negative basis functions. Hereby, we provide numerical examples and a summary that concludes the article.

2 Preliminaries

2.1 Linear Elliptic PDEs

A linear second order differential operator $L = \sum_{ij} a_{ij}\partial_{x_i}\partial_{x_j} + \sum_i b_i\partial_{x_i} + c$ with possibly spatially varying measurable coefficients a_{ij} (w.l.o.g. symmetric), b_i, c on a bounded domain $\Omega \subset \mathbb{R}^N$ is said to be strictly elliptic if there exists a constant $\lambda > 0$, such that

$$\sum_{ij} a_{ij}(x)\xi_i\xi_j \geq \lambda|\xi|^2 \text{ for every } \xi \in \mathbb{R}^N \text{ and a.e. } x \in \Omega\,. \tag{2}$$

In this article, we consider linear second-order differential operators

$$Lu := \operatorname{div}(a\nabla u) - cu \tag{3}$$

in divergence form, and require uniform ellipticity in the sense that the coefficients $a \in L^\infty(\Omega, \mathrm{Sym}(n \times n))$, $c \in L^\infty(\Omega, \mathbb{R})$ are at least bounded and the symmetric matrices $a = (a_{ij})$ are positive definite with smallest eigenvalue bounded away from zero on Ω by a constant $\lambda > 0$. Note that (3) can be rewritten under the additional assumption $a \in C^1(\Omega, \mathrm{Sym}(n \times n))$ in the above general form, and uniform ellipticity for bounded coefficients just means validity of (2). The corresponding Dirichlet boundary value problem reads as

$$-Lu = -\operatorname{div}(a\nabla u) + cu = f \text{ in } \Omega\,, \quad u = g \text{ on } \partial\Omega\,, \tag{4}$$

where we assume that the right-hand side (r.h.s., or inhomogeneity) satisfies at least $f \in (H_0^1(\Omega))^*$ and the boundary data satisfies $g \in H^{1/2}(\partial\Omega)$. In the case $c \geq 0$, the bilinear form

$$B(u, v) := \int_\Omega (a\nabla u) \cdot \nabla v + cuv\, dx \tag{5}$$

induced by $-L$ on the Sobolev space $H_0^1(\Omega)$ is coercive in the sense that $B(u, u) \geq \lambda\|\nabla u\|_2^2$ and bounded due to $|B(u, v)| \leq (\|a\|_\infty + C^2\|c\|_\infty)\|\nabla u\|_2\|\nabla v\|_2$ with the constant C in Sobolev's inequality $\|u\|_2 \leq C\|\nabla u\|_2$ for $u \in H_0^1(\Omega)$. Hence, by Lax–Milgram, (4) has a unique solution $u \in H^1(\Omega)$ satisfying the Dirichlet boundary condition in the sense that $u - g \in H_0^1(\Omega)$ for an extension of g from trace space $H^{1/2}(\partial\Omega)$ to $H^1(\Omega)$.

2.2 *Positivity, Maximum and Comparison Principle*

Definition 1 We say for a linear operator L on a space of functions on Ω that

- weak positivity holds if the validity of $-Lu \geq 0$ in Ω and $u \geq 0$ on $\partial\Omega$ implies the non-negativity $u \geq 0$ in Ω (resp. we say that strong positivity holds, if either $u \equiv 0$ or $u > 0$ in Ω is implied),
- the weak maximum principle holds if the validity of $Lu \geq 0$ in Ω implies that a non-negative maximum is attained by u on the boundary $\partial\Omega$ (resp. we say that the strong maximum principle holds, if either u is constant equal to its maximum or a non-negative maximum is attained by u only on the boundary $\partial\Omega$ and not inside Ω), or equivalently the weak minimum principle holds if the validity of $-Lu \geq 0$ in Ω implies that a non-positive minimum is attained by u on the boundary $\partial\Omega$ (resp. we say that the strong minimum principle holds, if either u is constant equal to its minimum or a non-positive minimum is attained by u only on the boundary $\partial\Omega$ and not inside Ω),

– the weak comparison principle holds if the validity of $-Lu_i = f_i$ in Ω, $u_i = g_i$ on $\partial\Omega$, $i = 1, 2$, implies for data $f_1 \geq f_2$, $g_1 \geq g_2$, that $u_1 \geq u_2$ holds in Ω (resp. the strong comparison principle holds, if either $u_1 \equiv u_2$ or $u_1 > u_2$ in Ω is implied).

To obtain the equivalence of minimum principle and maximum principle claimed in this definition, just substitute $-u$ for u. Note that Definition 1 is not fully precise, because no function space for u is provided. If $u \in C^2(\Omega) \cap C(\overline{\Omega})$ is assumed, then Definition 1 is called positivity, maximum principle or comparison principle for classical solutions. Positivity, maximum principle or comparison principle for weak solutions $u \in H^1(\Omega)$ requires a more precise definition indicated at the end of this subsection.

The weak minimum principle (and thus also the weak maximum principle) implies weak positivity (for classical solutions): If $-Lu \geq 0$ in Ω and $u \geq 0$ on $\partial\Omega$, then u cannot be negative in Ω, because by the weak minimum principle, a non-positive minimum would be attained on $\partial\Omega$ in contradiction to $u \geq 0$ on $\partial\Omega$, and hence $u \geq 0$ in Ω. Similarly, the strong minimum (or maximum) principle implies strong positivity.

Denote by $u = u_+ - u_-$ the decomposition of a function u into its positive part $u_+ := \max(u, 0) \geq 0$, and its negative part $u_- := \max(-u, 0) \geq 0$. For the convenience of the reader, we provide here proof of two well-known facts.

Lemma 1 *Every uniformly elliptic linear second-order differential operator L in divergence form (3) with $c \geq 0$ satisfies weak positivity (of weak solutions).*

Proof Assume that $-Lu = f \geq 0$ in Ω and $u \geq 0$ on $\partial\Omega$. Test $-Lu = f$ by u_- (note that $u \geq 0$ on $\partial\Omega$ implies $u_- = 0$ on $\partial\Omega$, thus u_- can act as a test function) and use $c \geq 0$ to obtain $\lambda\|u_-\|_{L^2}^2 \leq \int_\Omega (a\nabla u_-) \cdot \nabla u_- + c\,|u_-|^2\,dx = -\int_\Omega (a\nabla u) \cdot \nabla u_-\,dx - \int_\Omega c\,uu_-\,dx = \langle Lu, u_-\rangle = -\langle f, u_-\rangle \leq 0$ (because $f, u_- \geq 0$), i.e. $u_- \equiv 0$ and thus $u = u_+ \geq 0$ in Ω.

Remark 1 Weak positivity still holds for slightly negative c, as long as c is larger than the negative $-\lambda_1$ of the smallest Dirichlet eigenvalue λ_1 of $-L$.

Lemma 2 *For the differential operator L given by (3) with $c \geq 0$, weak positivity implies the weak maximum principle (for classical solutions).*

Proof Let L satisfy weak positivity, and let u be such that $Lu \geq 0$ and the maximum M of u on $\partial\Omega$ is non-negative, i.e. $M \geq 0$. Then $-L(M - u) \geq Lu \geq 0$ in Ω (as $c \geq 0$) and $M - u \geq 0$ on $\partial\Omega$. Thus, by weak positivity of L we have $M - u \geq 0$ in Ω and hence $u \leq M$ in Ω, i.e. a non-negative maximum of u is attained on the boundary $\partial\Omega$.

Similarly, for L given by (3) with $c \geq 0$, strong positivity implies the strong maximum principle (for classical solutions), and the weak (strong) comparison principle is equivalent to weak (strong) positivity, just put $u := u_1 - u_2$.

Yet, to show the strong maximum principle (or strong positivity) for weak solutions is more demanding. In its precise form, inequalities $f \geq 0$ for functionals $f \in (H_0^1(\Omega))^*$ have to be interpreted in the functional sense that $\langle f, v\rangle \geq 0$ for every $v \in H_0^1(\Omega)$ with $v \geq 0$ a.e. in Ω, and maximum / minimum have to be replaced by essential supremum/infimum. The strong maximum principle then states for a uniformly elliptic linear second order differential operator L in divergence form (3) with $c \geq 0$ (or even $c > -\lambda_1$) that $Lu \geq 0$ and $\sup_B u = \sup_\Omega u \geq 0$ for some closed ball B with positive radius in Ω imply $u \equiv \sup_\Omega u$ a.e. constant in Ω. For a proof of this strong maximum principle for weak solutions, the weak Harnack inequality can be applied to show that if $\sup_B u = \sup_\Omega u =: M \geq 0$ for some closed ball B with positive radius r in Ω, then $u \equiv M$ is constant on an even larger ball in Ω with radius greater than r, and a covering argument then allows to conclude $u \equiv M$ a.e. in Ω, see, e.g. [7, Theorem 8.19].

2.3 Bernstein Polynomials and Their Duals

As we aim to discuss in this article the Bernstein dual Petrov–Galerkin method for the approximation of a solution of (4), we need to discuss Bernstein polynomials and their dual polynomials over the N-dimensional unit cube $(0, 1)^N$. To formulate the Bernstein expansion of a real polynomial N-variate function, we use component-wise comparisons and arithmetic operations on *multiindices* $i = (i_1, \ldots, i_n) \in \mathbb{N}_0^N$. For $x \in \mathbb{R}^N$ and a multiindex $i \in \mathbb{N}_0^N$, its monomial is $x^i := x_1^{i_1} \ldots x_N^{i_N}$. Using compact notation $D = (D_1, \ldots, D_N) \in \mathbb{N}_0^N$, we put $\sum_{i=0}^{D} := \sum_{i_1=0}^{D_1} \ldots \sum_{i_N=0}^{D_N}$ and $\binom{D}{i} := \prod_{\mu=1}^{N} \binom{D_\mu}{i_\mu}$. An N-variate polynomial function u is expressed in *monomial form* as

$$u(x) = \sum_{i=0}^{d} a_i x^i, \tag{6}$$

where $d = (d_1, \ldots, d_n)$, and can be represented in *Bernstein form* by

$$u(x) = \sum_{j=0}^{D} u_j^{(D)} \, S_j^{(D)}(x), \quad x \in (0, 1)^N. \tag{7}$$

In (7), the jth Bernstein polynomial of degree $D \geq d$ is

$$S_j^{(D)}(x) = \binom{D}{j} x^j (1 - x)^{D-j}, \; x \in (0, 1)^N \tag{8}$$

and can be considered as tensor product of univariate Bernstein polynomials, i.e. $S_j^{(D)}(x) = S_{j_1}^{D_1}(x_1) \cdot \ldots \cdot S_{j_N}^{D_N}(x_N)$. Moreover, the *Bernstein coefficients* $u_j^{(D)}$ of degree D are given analytically in terms of the coefficients a_i in (6) by the formula

$$u_j^{(D)} = \sum_{i=0}^{j} \frac{\binom{j}{i}}{\binom{D}{i}} a_i, \ 0 \le j \le D. \tag{9}$$

Conversely, the following theorem from the literature provides a way of converting a polynomial from the Bernstein form to the monomial form.

Theorem 1 ([8, Theorem 3.3]) *Let $u(x)$ be a polynomial in Bernstein form of any degree D. Then its monomial form is*

$$u(x) = \sum_{i=0}^{D} a_i x^i,$$

where

$$a_i = \sum_{j=0}^{i} (-1)^{i-j} \binom{D}{i} \binom{i}{j} u_j^{(D)}, \ \ 0 \le i \le D.$$

We highlight two important properties of Bernstein polynomials, namely, the *end-point interpolation property*

$$u_j^{(D)} = u\left(\frac{j}{D}\right),$$

for $0 \le j \le D$ with $j_k \in \{0, D_k\}$, $k = 1, \ldots, N$, and the *enclosing property* [9]

$$\min_{0 \le j \le D} u_j^{(D)} \le u(x) \le \max_{0 \le j \le D} u_j^{(D)},$$

for all $x \in (0, 1)^N$. The parameters $D = (D_1, \ldots, D_N) \in \mathbb{N}_0^N$ determine in the mesh-free Bernstein dual Petrov–Galerkin method the resolution of the approximation in each coordinate direction, in analogy as the number of subdivisions of each interval in $(0, 1)^N$ determines how fine a rectangular mesh is in a FE method. In the following, for the convenience of the reader, we usually suppress the upper index containing the fixed degree D and mention it only, where it is helpful for understanding.

The dual polynomials to (one-dimensional) Bernstein polynomials in $L^2(0, 1)$ have been introduced by [10], who found a recurrence relation involving Legendre polynomials. We denote by $\tilde{\Psi}_i^{(D)}$ the N-variate dual Bernstein polynomials of degree $D \in \mathbb{N}_0^N$ determined by biorthogonality

$$\int_{(0,1)^N} S_j^{(D)} \tilde{\Psi}_i^{(D)} \, d\vec{x} = \delta_{ij} \, . \tag{10}$$

The coefficients c_{ij} in the decomposition $\tilde{\Psi}_i = \sum_{j=0}^{D} c_{ij} S_j$ are explicitly known (in one dimension) due to [11], see also [4, (2.4), (2.5)]. Here, we focus on linear combinations Ψ_i, $1 \le i \le D-1$, of the dual Bernstein polynomials called modal basis functions, which vanish on the boundary of the unit cube $(0,1)^N$. In contrast to [4, Proposition 1], we use an index shifted by one and a scaling so that relation [4, (2.8)] becomes

$$\Psi_i(x) := \tilde{a}_i \tilde{\Psi}_{i-1}(x) + \tilde{\Psi}_i(x) + \tilde{b}_i \tilde{\Psi}_{i+1}(x) \tag{11}$$

for $1 \le i \le D-1$ with $\tilde{a}_i = \frac{D-i+2}{2(i+1)}$, $\tilde{b}_i = \frac{i+2}{2(D-i+1)}$. Particularly, the space $\Pi_0^{(D)}$ of polynomial functions of maximal degree $D \in \mathbb{N}_0^N$ in each coordinate vanishing on the boundary of $(0,1)^N$ is not only spanned by S_j, $1 \le j \le D-1$, but also by Ψ_i, $1 \le i \le D-1$.

2.4 Certificates of Non-negativity Resp. Positivity

In a broad sense, every identity that gives immediate proof of non-negativity resp. positivity for a (multivariate) real function u is considered to be a certificate of non-negativity resp. positivity, and thus there are many different certificates of non-negativity resp. positivity. For example, a sum of squares (SOS) certificate of non-negativity for a real polynomial function u on $\mathbb{R}^N$ is a representation $u = \sum_{k=1}^{m} p_k^2$ of u by sums of squares of polynomials $p_1, \ldots, p_m$ on $\mathbb{R}^N$. However, not every real polynomial function $u \ge 0$ can be decomposed into a sum of squares of polynomials, e.g. the Motzkin polynomial $u(x,y) := x^4y^2 + x^2y^4 + 1 - 3x^2y^2$ on $\mathbb{R}^2$. On $\mathbb{R}_+^N = (0,\infty)^N$, a certificate of positivity for a real polynomial function u of degree $d \in \mathbb{N}_0^N$ in monomial form $u(x) = \sum_{i=0}^{d} a_i x^i$ is the validity of $a_i > 0$ for all $0 \le i \le d$. However, again not every real polynomial function $u > 0$ on $\mathbb{R}_+^N$ has positive monomial coefficients.

In this article, we consider Bernstein certificates: As the Bernstein basis polynomials $S_j^{(D)}$ in (8) are by construction non-negative on the closed unit cube $[0,1]^N$ and positive in its interior $(0,1)^N$, i.e. $S_j^{(D)}(x) > 0$ for all $x \in (0,1)^N$ and $0 \le j \le D$, where 0 denotes the multiindex with all components equal to zero, for a real polynomial function u on $\mathbb{R}^N$ the validity of $u_j^{(D)} \ge 0$ for all $0 \le j \le D$ implies non-negativity $u \ge 0$ on $[0,1]^N$. Thus, the non-negativity of Bernstein coefficients is a certificate of non-negativity on $[0,1]^N$. Further, if additionally $u_j^{(D)} > 0$ for one $0 \le j \le D$, then $u > 0$ on $(0,1)^N$, hence $u_j^{(D)} \ge 0$ for all $0 \le j \le D$ and $u_j^{(D)} \ne 0$ is a certificate of positivity on $(0,1)^N$. However, again there exist positive polynomials over a box which have non-positive Bernstein coefficients, as shown in the following example.

Example 1 Consider the polynomial $u(x) = 7x^2 - 3x + 5$. It is immediate to check that u is positive on $[-1, 1]$, but the list of Bernstein coefficients $(u_j^{(2)}) = (15, -2, 9)$ has a negative value. However, the polynomial u has a certificate of positivity at degree 3 on $[-1, 1]$, since $(u_j^{(3)})) = (15, 3.6, 1.6, 9)$.

3 Bernstein Dual Petrov–Galerkin Method

In using the Bernstein dual Petrov–Galerkin method [4] for solving (4) on a simply connected bounded domain $\Omega \subset \mathbb{R}^N$, the first step is to fix the degree $D = (D_1, \ldots, D_N) \in \mathbb{N}_0^N$ as a parameter which determines the resolution of the approximation in each coordinate, and a diffeomorphism $T : \Omega \to (0, 1)^N$ which extends continuously to a homeomorphism $T : \overline{\Omega} \to [0, 1]^N$. As basis functions, then the transformed Bernstein polynomials $S_j^{(D)}(T(x))$ are used (where in the following we usually suppress the upper index containing the fixed degree D), and particularly we search for an approximation of the solution of the form

$$u(x) = g(x) + \sum_{j=1}^{D-1} u_j S_j(T(x)) \,. \tag{12}$$

with an extension g of the boundary data to $H^1(\Omega)$. Note that the sum vanishes on the boundary $\partial\Omega$ due to $1 \le j \le D - 1$, where 1 denotes the multiindex with all components equal to one. Putting the ansatz (12) into the weak formulation $B(u, v) = \langle f, v\rangle$ with the bilinear form B from (5) and using as test functions v the transformed modal basis functions $\Psi_i \circ T$ vanishing on the boundary of the unit cube $(0, 1)^N$, which are induced via (11) by the dual Bernstein polynomials $\tilde{\Psi}_i$, we obtain a linear system

$$\mathbf{A}\vec{u} = \vec{b} \tag{13}$$

with stiffness matrix $\mathbf{A} = \left(\int_\Omega (a\nabla(S_j \circ T)) \cdot \nabla(\Psi_i \circ T) + c(S_j \circ T)(\Psi_i \circ T)\, d\vec{x}\right)$ and r.h.s. $\vec{b} = (\langle f, \Psi_i \circ T\rangle - B(g, \Psi_i \circ T))$.

Example 2 If $\Omega := (\underline{x}_1, \overline{x}_1) \times \cdots \times (\underline{x}_N, \overline{x}_N)$ is a general open n-dimensional cube with

$$\underline{x}_\mu < \overline{x}_\mu, \ \ \mu = 1, \ldots, N$$

identified by the usual affine linear transform $T : \Omega \to (0, 1)^N$ with the unit cube, then the transformed jth Bernstein polynomial of degree $D \in \mathbb{N}_0^N$ is

$$S_j^{(D)}(T(x)) = \binom{D}{i} (x - \underline{x})^i (\overline{x} - x)^{D-i} w(\Omega)^{-D}, \tag{14}$$

where $w(\Omega) = (\overline{x}_1 - \underline{x}_1, \ldots, \overline{x}_N - \underline{x}_N)$ denotes the width of intervals.

Due to the chain rule and transformation formula, system (13) is identical with the system obtained from (4) on the unit cube $(0,1)^N$ with coefficients replaced by $\left(\frac{1}{|\det DT|}(DT)a(DT)^*\right)\circ T^{-1}$ resp. $\left(\frac{1}{|\det DT|}c\right)\circ T^{-1}$. Note that this transformation does not change the uniform ellipticity and boundedness of the coefficients. Therefore, we can restrict our discussion to the case of the unit cube.

In the special case, where after the transformation the coefficients of (4) on the unit cube $(0,1)^N$ are given by constants $a = \mathbf{I}$ equal to the identity matrix $\mathbf{I}$ and $c = 0$, and the boundary value $g = 0$ vanishes, we are in case of Poisson's equation (1) on the unit cube $(0,1)^N$ subject to homogeneous Dirichlet boundary conditions. For this case, in dimension $N = 2$ it can be seen from [4, 3.2.1] that the stiffness matrix $\mathbf{A}$ in (13) can be written as tensor product $\mathbf{A} = A \otimes B + B \otimes A$ with sparse matrices $A = \left(\int_0^1 S_j'(x)\Psi_i'(x)\,dx\right)$, $B = \left(\int_0^1 S_j(x)\Psi_i(x)\,dx\right)$, containing one-dimensional integrals of univariate (dual) Bernstein polynomials and their first derivatives. Hereby, sparsity follows from the validity of the three-term recurrence relation [12], [4, (2.3)],

$$S_j'(x) = (D-j+1)S_{j-1}(x) - (D-2j)S_j(x) - (j+1)S_{j+1}(x) \tag{15}$$

for the one-dimensional derivative of Bernstein polynomials.

Similarly, [4, Corollary 1] offers a five-term recurrence relation for the derivative of one-dimensional dual Bernstein polynomials, and together with the three-term recurrence relation for the derivative of one-dimensional Bernstein polynomials and biorthogonality (10) sparsity of the one-dimensional matrices A, B and thus of the stiffness matrix $\mathbf{A} = A \otimes B + B \otimes A$ follows. Yet, even in case of Poisson's equation (1) on the unit cube $(0,1)^N$ subject to homogeneous Dirichlet boundary conditions, this stiffness matrix does not have a non-negative inverse.

Example 3 In the case $N = 2$, $D = (6,6)$, the one-dimensional matrices $A, B \in \mathbb{R}^{5,5}$ are given by

$$\begin{aligned} A &= \begin{pmatrix} 58+\frac{2}{7} & -10-\frac{2}{7} & -10-\frac{2}{7} & -1-\frac{5}{7} & 0 \\ -9 & 28.5 & -1.5 & -9 & -1.5 \\ -8.64 & -1.44 & 21.76 & -1.44 & -8.64 \\ -1.5 & -9 & -1.5 & 28.5 & -9 \\ 0 & -1-\frac{5}{7} & -10-\frac{2}{7} & -10-\frac{2}{7} & 58+\frac{2}{7} \end{pmatrix} \\ B &= \begin{pmatrix} 1 & 1 & 0 & 0 & 0 \\ \frac{5}{8} & 1 & \frac{5}{8} & 0 & 0 \\ 0 & 1 & 1 & 1 & 0 \\ 0 & 0 & \frac{7}{4} & 1 & \frac{1}{4} \\ 0 & 0 & 0 & 4 & 1 \end{pmatrix} \end{aligned} \tag{16}$$

Note that in contrast to [4] due to our scaling (11) here A has the symmetry $a_{6-i,6-j} = a_{i,j}$ for all $i, j \in \{1,2,3,4,5\}$ and B has values 1 on the diagonal. Further, while A^{-1} is positive, the stiffness matrix $\mathbf{A} = A \otimes B + B \otimes A \in \mathbb{R}^{25,25}$ has negative entries.

Remark 2 Moreover, in the general situation of an arbitrary domain and arbitrary coefficients, the stiffness matrix $\mathbf{A}$ is neither sparse nor has a non-negative inverse $\mathbf{A}^{-1}$.

While we will see in the next section that a non-negative inverse of the stiffness matrix $\mathbf{A}$ is related to algebraic positivity, for the formulation of the algebraic discrete maximum principle let us fix an approximation by Bernstein polynomials of the right- hand side $f(x) = \sum_{j=0}^{D} f_j S_j(T(x))$ and of the (extended) boundary data $g(x) = \sum_{j=0,D} g_j S_j(T(x))$, where $j = 0, D$ means that at least one index satisfies $j_k \in \{0, D_k\}$ for $k = 1, \ldots, N$, and let us extend the system (13) to

$$\begin{pmatrix} \mathbf{A} & \mathbf{A}^{\partial} \\ \mathbf{0} & \mathbf{I} \end{pmatrix} \begin{pmatrix} \vec{u} \\ \vec{g} \end{pmatrix} = \begin{pmatrix} \mathbf{M}\vec{f} \\ \vec{g} \end{pmatrix} \tag{17}$$

with $\mathbf{A}^{\partial} = \left(\int_{\Omega} (a\nabla(S_j \circ T)) \cdot \nabla(\Psi_i \circ T) + c\,(S_j \circ T)(\Psi_i \circ T)\, d\vec{x}\right)_{1 \le i \le D-1,\, j=0,D}$, and $\mathbf{M} := \left(\int_{\Omega} (S_j \circ T)(\Psi_i \circ T)\, d\vec{x}\right)_{1 \le i \le D-1,\, 0 \le j \le D}$. In contrast to (13), where the right hand side f and the boundary data g are hidden within the vector $\vec{b} = (\langle f, \Psi_i \circ T\rangle - B(g, \Psi_i \circ T))$, in (17) the dependence of the algebraic solution on the coefficients of the data is made explicit.

Remark 3 Note that the discussion of a discrete maximum principle by [2] is based on the extended linear system (17). Not only in case of FEM or the Bernstein dual Petrov–Galerkin method, but in an arbitrary Galerkin method also, such an extension can always be obtained by projecting boundary data on a finite dimensional space spanned by boundary basis functions.

4 Discrete Maximum Principle and Positivity Certificates

Discrete positivity and the discrete maximum/minimum principle can be valid in two different ways, algebraically or functionally. Let us begin our discussion by defining algebraic discrete positivity and the algebraic discrete maximum/minimum principle.

Definition 2 We say that the extended linear system (17) satisfies

- algebraic weak discrete positivity, if data $\vec{f}, \vec{g} \ge 0$ implies a solution $\vec{u} \ge 0$ (resp. algebraic strong discrete positivity, if either $(\vec{u}, \vec{g}) = 0$ or $\vec{u} > 0$ is implied),
- the algebraic weak discrete maximum principle, if data $\vec{f} \le 0$ implies that a non-negative maximal component of $(\vec{u}, \vec{g})$ already occurs in $\vec{g}$ (resp. the algebraic strong discrete maximum principle, if it is implied that either all components of $(\vec{u}, \vec{g})$ are identical or a non-negative maximal component of $(\vec{u}, \vec{g})$ occurs only in $\vec{g}$),

or equivalently the algebraic weak discrete minimum principle, i.e. data $\vec{f} \geq 0$ implies that a non-positive minimal value of $(\vec{u}, \vec{g})$ already occurs in $\vec{g}$ (resp. the algebraic strong discrete minimum, if it is implied that either all components of $(\vec{u}, \vec{g})$ are identical or a non-positive minimal component of $(\vec{u}, \vec{g})$ occurs only in $\vec{g}$)

The following discrete analogue of Lemma 2 allows to restrict our attention to algebraic discrete positivity.

Lemma 3 *If the matrix in* (17) *has non-negative row sums and* $\mathbf{M} \geq 0$*, then algebraic weak (resp. strong) discrete positivity implies the algebraic weak (resp. strong) discrete maximum principle.*

Proof Let (17) satisfy algebraic weak positivity, let $\vec{f} \leq 0$, and assume that the maximal value M in a component of $\vec{g}$ is non-negative. Then

$$\mathbf{A}(M\vec{1} - \vec{u}) + \mathbf{A}^{\partial}(M\vec{1} - \vec{g}) \geq -\mathbf{A}\vec{u} - \mathbf{A}^{\partial}\vec{g} = -\mathbf{M}\vec{f} \geq 0 \tag{18}$$

due to non-negative row sums $\mathbf{A}\vec{1} + \mathbf{A}^{\partial}\vec{1} \geq 0$ and non-negativity $\mathbf{M} \geq 0$, and $M\vec{1} - \vec{g} \geq 0$. Thus, by algebraic weak positivity of (17) we have $M\vec{1} - \vec{u} \geq 0$ and hence $\vec{u} \leq M\vec{1}$, i.e. a non-negative maximal component of $(\vec{u}, \vec{g})$ already occurs in $\vec{g}$. In case of algebraic strong positivity, we obtain in the last step of the former proof either $(M\vec{1} - \vec{u}, M\vec{1} - \vec{g}) = 0$ or $M\vec{1} - \vec{u} > 0$, i.e. either all components of $(\vec{u}, \vec{g})$ are identical with M or $\vec{u} < M\vec{1}$ so that a non-negative maximal component of $(\vec{u}, \vec{g})$ occurs only in $\vec{g}$.

The following Lemma allows to prove algebraic weak (resp. strong) positivity and hence the algebraic weak (resp. strong) discrete maximum principle for several Galerkin methods, e.g. piecewise linear FEM on simplices under the interior edge condition.

Lemma 4 *If the matrix in* (17) *has non-negative row sums, if* $\mathbf{A}$ *is a non-singular* M*-matrix,* $\mathbf{A}^{\partial} \leq 0$ *and* $\mathbf{M} \geq 0$*, then algebraic weak discrete positivity holds. If moreover, at least one-row sum is positive,* $\mathbf{A}$ *is irreducible and* $\mathbf{M}$ *is surjective, then even algebraic strong discrete positivity holds.*

Proof By [13], $\mathbf{A}$ is a non-singular M-matrix iff $\mathbf{A}^{-1}$ exists and is non-negative. Therefore, the matrix in (17) hat the inverse $\begin{pmatrix} \mathbf{A}^{-1} & -\mathbf{A}^{-1}\mathbf{A}^{\partial} \\ 0 & \mathbf{I} \end{pmatrix}$, which is non-negative due to $\mathbf{A}^{\partial} \leq 0$. Moreover, under the additional assumptions, $\mathbf{A}$ is an irreducibly diagonally dominant real square matrix with strictly positive diagonal and non-positive off-diagonal entries, and thus $\mathbf{A}^{-1} > 0$ by [14].

If the stiffness matrix $\mathbf{A}$ of the linear system (13) does not have a non-negative inverse $\mathbf{A}^{-1}$, then algebraic (weak) positivity is not valid. In fact, if $\mathbf{A}^{-1}$ has an element $a_{ij}^{-1} < 0$, then for $\vec{g} := \vec{0}$ the jth component of $\mathbf{M}\vec{f} \geq 0$ can be chosen so large that $\vec{u} = \mathbf{A}^{-1}\mathbf{M}\vec{f}$ has a negative ith component. This is the case for Bernstein

dual Petrov–Galerkin method by Example 3 and Remark 2. Yet, on the one hand, for some data, it is still possible to obtain a Bernstein certificate of non-negativity (resp. positivity) for the approximate solution.

Definition 3 For data $\vec{f}$ and $\vec{g} \geq 0$, we say that the extended linear system (17) allows a Bernstein certificate of non-negativity for the approximate solution u given by (12), if $\vec{u} \geq 0$ holds (resp. a Bernstein certificate of positivity, if additionally $\vec{u} \neq 0$ holds).

On the other hand, instead of algebraic weak discrete positivity $\vec{u} \geq 0$ it may be possible to conclude merely functional weak discrete positivity, where the weaker conclusion $u \geq 0$ for the approximate solution (12) with Bernstein coefficients $\vec{u}$ is drawn.

Definition 4 We say that the extended linear system (17) satisfies

- algebraic-functional weak discrete positivity, if data $\vec{f}, \vec{g} \geq 0$ implies $u \geq 0$ for the approximate solution (12) (resp. algebraic-functional strong discrete positivity, if either $u = 0$ or $u > 0$ is implied).
- functional-functional weak discrete positivity, if data $f, g \geq 0$ implies $u \geq 0$ for the approximate solution (12) (resp. functional-functional strong discrete positivity, if either $u = 0$ or $u > 0$ is implied).

Note that in this double notation, the first point in Definition 2 would be algebraic-algebraic discrete positivity (while for functional-algebraic weak discrete positivity the weakest conditions $f, g \geq 0$ would need to imply the strongest condition $\vec{u} \geq 0$). It is not astonishing that many spectral methods and particularly the Bernstein dual Petrov–Galerkin method do not satisfy algebraic discrete positivity, because signs of the coefficients $\langle f, \Psi_i \circ T \rangle$ do not say much about non-negativity resp. positivity of the function f, as the test functions Ψ_i itself are sign-changing. Therefore, functional discrete positivity is more important, however, also more difficult to verify, because the convex cone of non-negative functions is not finitely generated. To clarify, whether functional weak (resp. strong) positivity is valid, or for which data we can provide a Bernstein certificate of non-negativity (resp. positivity), let us apply the general theory of convex cones.

Definition 5 A subset $K \subset X$ of a real vector space X is called

- a cone, if $x \in K$ implies $rx \in K$ for every $r \geq 0$,
- convex, if $x, y \in K$ imply $\lambda x + (1 - \lambda) y \in K$ for every $\lambda \in [0, 1]$.

Thus, a convex cone $K \subset X$ is a subset such that $ax + by \in K$ for every linear combination with non-negative coefficients $a, b \geq 0$. If X is a Banach space, then a convex cone K is called closed if it is closed w.r.t. norm topology. For example, $\{\vec{u} \in \mathbb{R}^n \mid \vec{u} \geq 0\}$ is a closed convex cone in $\mathbb{R}^n$, and the subset $\{u \in H^1(\Omega) \mid u \geq 0 \text{ a.e.}\}$ is a closed convex cone in $H^1(\Omega)$.

Definition 6 The polar cone of a convex cone $K \subset X$ in a real Banach space is the subset $K^o := \{f \in X^* \mid \forall x \in K : \langle f, x \rangle \leq 0\}$ of the dual space X^*.

The polar of a convex cone is automatically closed in X^*, and for a closed convex cone $K \subset X$ in a reflexive Banach space $X \cong X^{**}$ the bipolar theorem $(K^o)^o = K$ holds. This fact allows the following characterization of data b such that the solution u of $Au = b$ lies in K.

Theorem 2 *Let $K \subset X$ be a closed convex cone in a reflexive Banach space X, and let $A : X \to X^*$ a bijective continuous linear map. Then the image $A(K)$ is a closed convex cone in X^*, and with the polar cone $P := (A(K))^o \subset X^{**} = X$ of $A(K)$ the following characterization holds:*

The unique solution $u \in X$ of $Au = b$ satisfies $u \in K$ iff $b \in P^o$.

Proof Eventually, $A(K)$ is a convex cone, and by the open mapping / closed graph theorem the image $A(K)$ is closed. By the bipolar theorem, $b \in P^o$ is equivalent to $b \in ((A(K))^o)^o = A(K)$, and thus $P^o \ni b = Au$ is equivalent to $u \in K$ by the uniqueness of solutions.

While the former using the bipolar theorem is more formal, the Theorem shows that those data b, for which $u \in K$ can be concluded, form a closed convex cone (namely P^o). As a consequence, the following results about positivity certificates and functional discrete positivity are valid.

Corollary 1 *Precisely for data $\vec{f}$ and $\vec{g} \geq 0$ satisfying $\mathbf{A}^{-1}\mathbf{M}\vec{f} - \mathbf{A}^{-1}\mathbf{A}^{\partial}\vec{g} \geq 0$ (resp. additionally a strict inequality > 0 in at least one component) a Bernstein certificate of non-negativity (resp. positivity) for the approximate solution can be provided.*

Proof Let $K := \{\begin{pmatrix}\vec{u}\\ \vec{g}\end{pmatrix} \mid \vec{u} \geq 0,\ \vec{g} \geq 0\}$, then for $\vec{g} \geq 0$ we have $\begin{pmatrix}\mathbf{M}\vec{f}\\ \vec{g}\end{pmatrix} \in P^o = \begin{pmatrix}\mathbf{A} & \mathbf{A}^{\partial}\\ \mathbf{0} & \mathbf{I}\end{pmatrix}(K)$ iff $\mathbf{A}^{-1}\mathbf{M}\vec{f} - \mathbf{A}^{-1}\mathbf{A}^{\partial}\vec{g} \geq 0$ (and if additionally a strict inequality > 0 holds in at least one component, then $\vec{u} \neq 0$ and thus the approximate solution u is positive).

Although this precise characterization involves the inverse, the inequalities in Corollary 1 define a convex cone, which can be practically used to verify positivity of an approximate solution without solving the linear system (17).

Example 4 In the special case of Poisson's problem (1) on the unit cube $(0, 1)^N$ with homogeneous Dirichlet boundary conditions, $N = 2$, $D = (6, 6)$, and with the matrices $A, B \in \mathbb{R}^{5,5}$ provided in Example 3, the stiffness matrix is given by $\mathbf{A} = A \otimes B + B \otimes A \in \mathbb{R}^{25,25}$. Its inverse (for better readability scaled and rounded) $100\mathbf{A}^{-1}$ reads as

$$
\begin{pmatrix}
0.9 & -0.2 & -0.5 & 0.5 & -0.1 & -0.2 & 0.2 & -0 & -0.2 & 0.1 & -0.5 & -0 & 1.1 & -0.5 & 0 & 0.5 & -0.2 & -0.5 & 0.4 & -0 & -0.1 & 0.1 & 0 & -0 & -0 \\
-0.1 & 0.5 & 1 & -0.8 & 0.2 & 0.2 & -0 & -0.7 & 0.7 & -0.1 & -0 & 0.8 & -0.8 & 0.5 & -0.1 & -0.1 & -0.1 & 0.4 & -0.3 & 0.1 & 0.1 & -0.1 & 0.1 & -0 & 0 \\
-0.4 & 1.5 & -1.3 & 1.1 & -0.1 & -0.2 & -1 & 2 & -0.9 & 0.1 & 1.2 & -1.2 & 1 & -0.6 & 0.2 & -0.6 & 0.6 & -0.3 & 0.3 & -0.1 & 0 & 0.2 & -0.2 & 0.1 & -0.0 \\
1.6 & -3.6 & 2.9 & -0.5 & 0.2 & -1.3 & 2.9 & -2 & 0.6 & -0.1 & -0.7 & 1.9 & -2.2 & 1.2 & -0.1 & 0.7 & -1.4 & 1.3 & -0.4 & 0 & -0 & 0 & 0.2 & -0.1 & 0.0 \\
-5.5 & 10.1 & -6.9 & 1.3 & 0.4 & 6.9 & -7.9 & 2.3 & 0.1 & -0 & -1.8 & -4.8 & 8.4 & -2.7 & 0.1 & -0.4 & 4.6 & -5 & 1 & 0.2 & 0 & -0.4 & 0.3 & 0.2 & -0.1 \\
-0.1 & 0.2 & -0 & -0.1 & 0.1 & 0.5 & -0 & 0.8 & -0.1 & -0.1 & 1 & -0.7 & -0.8 & 0.4 & 0.1 & -0.8 & 0.7 & 0.5 & -0.3 & -0 & 0.2 & -0.1 & -0.1 & 0.1 & 0.0 \\
0.1 & -0 & -0.4 & 0.4 & -0 & -0 & 1.7 & 0.2 & -0.2 & 0.1 & -0.4 & 0.2 & -0.6 & 0.4 & 0 & 0.4 & -0.2 & 0.4 & -0.1 & -0 & -0 & 0.1 & 0 & -0 & 0.0 \\
-0.1 & -0.5 & 1.2 & -0.5 & 0.1 & 1 & 0.3 & -0.2 & 0.5 & 0.1 & -1.2 & -0.8 & 2.4 & -0.9 & -0 & 0.7 & 0.5 & -1.2 & 0.5 & 0.1 & -0.1 & 0 & 0.1 & 0 & -0.0 \\
-0.8 & 1.7 & -1.1 & 0.4 & -0 & 0.9 & -1.7 & 0.6 & 1.1 & 0.2 & 0 & 0.8 & -0.8 & 0.4 & -0.2 & -0.4 & 0.3 & 0.3 & -0.2 & 0.2 & 0.1 & -0.2 & 0.2 & 0 & -0.0 \\
4.2 & -4.8 & 1.3 & 0.2 & -0 & -8.1 & 5 & 2.7 & -1.6 & 0.6 & 6.1 & 0.9 & -6.1 & 0.6 & 0.6 & -1.8 & -2.9 & 4.3 & 0.1 & -0.4 & 0.1 & 0.9 & -1 & 0.1 & 0.1 \\
-0.4 & -0.2 & 1.2 & -0.6 & 0 & 1.5 & -1 & -1.2 & 0.6 & 0.2 & -1.3 & 2 & 0.9 & -0.3 & -0.2 & 1.1 & -0.9 & -0.6 & 0.3 & 0.1 & -0.1 & 0.1 & 0.2 & -0.1 & -0.0 \\
-0.1 & 1 & -1.2 & 0.7 & -0.1 & -0.5 & 0.3 & -0.8 & 0.5 & 0 & 1.2 & -0.2 & 2.4 & -1.2 & 0.1 & -0.5 & 0.5 & -0.9 & 0.5 & 0 & 0.1 & 0.1 & -0 & 0.1 & -0.0 \\
1.2 & -1.7 & 1.6 & -1 & 0.3 & -1.7 & -1.2 & 3.5 & -1.3 & -0 & 1.6 & 3.5 & -5.4 & 2.9 & -0 & -1 & -1.3 & 2.9 & -1.1 & 0 & 0.3 & -0 & -0 & 0 & 0.1 \\
-0.9 & 2.5 & -3 & 1.6 & -0.2 & 0 & 1.1 & -1.2 & 0.6 & -0.3 & 0.7 & -3.8 & 4.5 & -1.3 & 0.8 & 0 & 1.1 & -1.4 & 0.8 & -0.3 & -0.1 & 0.2 & -0.2 & 0.2 & 0.0 \\
-1 & -5.7 & 9.2 & -3.4 & 0.2 & 8.7 & 1.4 & -8.8 & 0.8 & 0.8 & -12 & 3 & 4.8 & 2.2 & -0.9 & 5.3 & 0.9 & -4.7 & -0.4 & 0.7 & -0.6 & -1 & 1.6 & -0.3 & -0.0 \\
1.6 & -1.3 & -0.7 & 0.7 & -0.1 & -3.7 & 2.9 & 1.9 & -1.4 & 0 & 2.9 & -2 & -2.2 & 1.3 & 0.2 & -0.5 & 0.6 & 1.2 & -0.4 & -0.1 & 0.2 & -0.1 & -0.1 & 0 & 0.0 \\
-0.8 & 0.9 & 0 & -0.4 & 0.1 & 1.7 & -1.7 & 0.8 & 0.3 & -0.2 & -1.1 & 0.6 & -0.8 & 0.3 & 0.2 & 0.4 & 1.1 & 0.4 & -0.2 & 0 & -0 & 0.2 & -0.2 & 0.2 & -0.0 \\
-0.9 & 0 & 0.7 & 0 & -0.1 & 2.5 & 1.1 & -3.8 & 1.1 & 0.2 & -3 & -1.2 & 4.5 & -1.4 & -0.2 & 1.6 & 0.6 & -1.3 & 0.8 & 0.2 & -0.2 & -0.3 & 0.7 & -0.3 & 0.0 \\
1.8 & -2.8 & 1.3 & 0.2 & -0.1 & -2.8 & 3.7 & -0.9 & -0.8 & 0.5 & 1.3 & -0.9 & -0.1 & 0.4 & -0.4 & 0.2 & -0.8 & 0.4 & 0.7 & 0.3 & -0.1 & 0.5 & -0.4 & 0.3 & -0.0 \\
-3.3 & 12 & -9 & -0.3 & 1 & -0.6 & -20 & 20.2 & 0.4 & -2.2 & 9.4 & 10.2 & -18.8 & 1.6 & 1.6 & -7.9 & -0 & 7.4 & -1.4 & 0 & 1.6 & -0.7 & -0.6 & 0 & 0.2 \\
-5.5 & 6.9 & -1.8 & -0.4 & 0 & 10.1 & -7.9 & -4.8 & 4.6 & -0.4 & -6.9 & 2.3 & 8.4 & -5 & 0.3 & 1.3 & 0.1 & -2.7 & 1 & 0.2 & 0.4 & -0 & 0.1 & 0.2 & -0.1 \\
4.2 & -8.1 & 6.1 & -1.8 & 0.1 & -4.8 & 5 & 0.9 & -2.9 & 0.9 & 1.3 & 2.7 & -6.1 & 4.3 & -1 & 0.2 & -1.6 & 0.6 & 0.1 & 0.1 & -0 & 0.6 & 0.6 & -0.4 & 0.1 \\
-1 & 8.7 & -12 & 5.3 & -0.6 & -5.7 & 1.4 & 3 & 0.9 & -1 & 9.2 & -8.8 & 4.8 & -4.7 & 1.6 & -3.3 & 0.8 & 2.2 & -0.4 & -0.3 & 0.2 & 0.8 & -0.9 & 0.7 & -0.0 \\
-3.3 & -0.6 & 9.4 & -7.9 & 1.6 & 12 & -19.9 & 10.2 & -0 & -0.7 & -9 & 20.2 & -18.8 & 7.4 & -0.6 & -0.3 & 0.4 & 1.6 & -1.4 & 0 & 1 & -2.2 & 1.6 & 0 & 0.2 \\
13.1 & -22.7 & -0.8 & 16.6 & -5.2 & -22.7 & 79.8 & -60.8 & -0.5 & 5.9 & -0.8 & -60.8 & 75.6 & -17.9 & -1.9 & 16.6 & -0.5 & -17.9 & 6.5 & 0.1 & -5.2 & 5.9 & -1.9 & 0.1 & 0.3
\end{pmatrix}
$$

and obviously is not non-negative. The mass matrix $\mathbf{M} \geq 0$ is given by the extension of $B \otimes B$ to a (25×49)-matrix containing additionally the values of $\int_0^1 S_j(x)\Psi_i(x)\,dx$ for $1 \leq i \leq D-1$, $j = 0, D$, and again $\mathbf{A}^{-1}\mathbf{M} \in \mathbb{R}^{25\times 49}$ is not non-negative. Thus, by Corollary 1 for Poisson's problem (1) on the unit cube $(0, 1)^2$ with boundary data $g = 0$, a Bernstein certificate of non-negativity for the approximate solution u can be provided for r.h.s. f with Bernstein coefficients $\vec{f}$ in the cone C given by inequalities $\mathbf{A}^{-1}\mathbf{M}\vec{f} \geq 0$. Due to missing non-negativity of $\mathbf{A}^{-1}\mathbf{M}$, the cone $\{\vec{f} \mid \vec{f} \geq 0\}$ is not a subset of C, and while e.g. $\mathbb{R}^{25+24} \ni \vec{f} = (1, 0, \ldots, 0)^T \notin C$, the Bernstein coefficients $\mathbb{R}^{25+24} \ni \vec{f} = (1, \ldots, 1, 0, \ldots, 0)^T$ satisfy $\mathbf{A}^{-1}\mathbf{M}\vec{f} \geq 0$ and thus lie in C. Hence, $\vec{f} = (1, \ldots, 1, 0, \ldots, 0)^T$ (and $\vec{g} = 0$) allow a Bernstein certificate of non-negativity for the approximate solution u, precisely the (scaled) Bernstein coefficients of u are given by

$$10\vec{u} = (0.41, 0.67, 0.76, 0.71, 0.57, 0.67, 1.15, 1.29, 1.23, 0.87, 0.76, 1.29, 1.45, 1.41, 1.04, 0.71, 1.23, 1.41, 1.32, 0.91, 0.57, 0.87, 1.04, 0.91, 0.97)^T .$$

Of course, it would be nice to have more simple inequalities for $\hat{f}$ (and $\hat{g} \geq 0$) or equivalently more simple closed convex cones, which guarantee a certificate of non-negativity $\vec{u} \geq 0$. Obtaining such simplification strongly depends on the chosen method and will be a task for a forthcoming paper about the Bernstein dual Petrov–Galerkin method.

Note that the cone $K = \{\begin{pmatrix}\vec{u}\\ \vec{g}\end{pmatrix} \mid \vec{u} \geq 0,\ \vec{g} \geq 0\} = \mathrm{cone}(\{\vec{e}_j \mid 0 \leq j \leq D\})$ in the former proof is finitely generated by the unit vectors. This is not the case in the next Corollary characterizing functional weak positivity, what makes it more difficult to apply the Corollary.

Corollary 2 *Denote by* $K := \{\begin{pmatrix}\vec{u}\\ \vec{g}\end{pmatrix} \mid \forall x \in \Omega : \sum_{j=0,D} g_j S_j(T(x)) + \sum_{j=1}^{D-1} u_j S_j(T(x)) \geq 0\}$ *the convex cone of Bernstein coefficients of non-negative Bernstein polynomials of degree D, then algebraic-functional weak discrete positivity holds iff the*

matrix $\begin{pmatrix} \mathbf{A}^{-1}\mathbf{M} & -\mathbf{A}^{-1}\mathbf{A}^{\partial} \\ 0 & \mathbf{I} \end{pmatrix}$ *maps the convex cone* $C_1 := \{\begin{pmatrix} \vec{f} \\ \vec{g} \end{pmatrix} \mid \vec{f}, \vec{g} \geq 0\}$ *into* K, *and functional-functional weak discrete positivity holds iff the convex cone*

$$C_2 := \{\begin{pmatrix} \vec{f} \\ \vec{g} \end{pmatrix} \mid \forall x \in \Omega : \sum_{j=0,D} g_j S_j(T(x)) \geq 0, \sum_{j=0}^{D} f_j S_j(T(x)) \geq 0\}$$

is mapped into K.

Similarly, algebraic-functional or functional-functional strong discrete positivity can be characterized, using cones K without zero $K \setminus \{0\}$ or the pointed interior cone $\overset{o}{K} \cup \{0\}$. Hereby, in the infinite-dimensional case it is important to work in $H^1(\Omega)$ or a stronger space, because else even the standard cone does not have a non-empty interior.

Remark 4 While the convex cone $\{u \in L^2(\Omega) \mid u \geq 0 \text{ a.e.}\}$ is closed in $L^2(\Omega)$, it has no interior points. In fact, a function $u \in L^2(\Omega)$ satisfying $u > 0$ a.e. and $u < M$ a.e. on $B_{\epsilon_0}(x_0)$ can be perturbed by subtracting $M 1_{B_\epsilon(x_0)}$ for sufficiently small $\epsilon < \epsilon_0$. The resulting function $u - M 1_{B_\epsilon(x_0)}$ is negative on $B_\epsilon(x_0)$, although the norm $\|M 1_{B_\epsilon(x_0)}\|_{L^2} = M \operatorname{Vol}(B_\epsilon(x_0))$ is arbitrarily small as $\epsilon \searrow 0$.

Let us conclude this section with a discussion of algebraic-functional weak discrete positivity in our main example.

Example 5 In the special case of Poisson's problem (1) on the unit cube $(0, 1)^N$ with homogeneous Dirichlet boundary conditions, $N = 2$, $D = (6, 6)$, already discussed

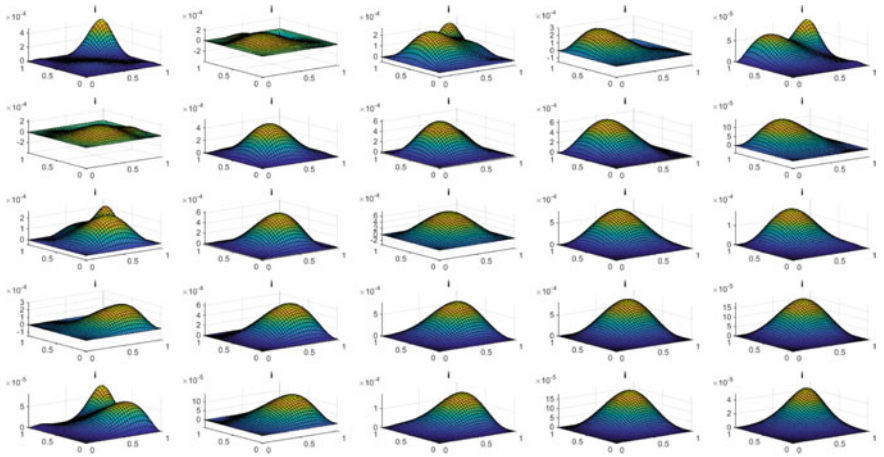

Fig. 1 The 25 approximate solutions of Poisson's problem (1) on the unit cube $(0, 1)^2$ obtained by Bernstein dual Petrov–Galerkin method for $D = (6, 6)$ and Bernstein polynomials $f = S_i$, $1 \leq i \leq D - 1$, vanishing on $\partial(0, 1)^2$ as r.h.s

in Examples 3 and 4, Fig. 1 shows the 25 approximate solutions u_i to those Bernstein polynomials $f = S_i$, $1 \leq i \leq D-1$ as r.h.s. which vanish on $\partial(0,1)^N$, i.e. those corresponding to $\vec{g} = 0$ and unit vectors $\vec{f} = \vec{e}_i$. As not all of these approximate solutions are non-negative, the cone C_1 from Corollary (2) is not mapped into K, i.e. algebraic-functional weak discrete positivity (and thus also the stronger functional-functional weak discrete positivity) does not hold for the Bernstein dual Petrov–Galerkin method. However, as merely the functions $u_{(1,2)}$, $u_{(2,1)}$, $u_{(1,3)}$, $u_{(3,1)}$, $u_{(1,4)}$, $u_{(4,1)}$, $u_{(2,5)}$, $u_{(5,2)}$ and $u_{(3,3)}$ become negative, we can guarantee non-negativity of approximate solutions u to r.h.s. having Bernstein coefficients $\vec{f}$ with vanishing components at indices (1, 2), (2, 1), (1, 3), (3, 1), (1, 4), (4, 1), (2, 5), (5, 2), (3, 3), even although the Bernstein coefficients $\vec{u}$ given by the columns of the matrix $\mathbf{A}$ in Example 4 are not non-negative for other indices, too.

5 Conclusion

In this article, we have completely characterized those data for which a Bernstein certificate of non-negativity (resp. positivity) can be given for the approximate solution of an elliptic linear second-order PDEs in divergence form when using Bernstein dual Petrov–Galerkin method. Further, we provided necessary and sufficient conditions for the validity of algebraic-functional or functional-functional discrete positivity, or equivalently for validity of the corresponding discrete maximum principles. Our methods can be directly transferred to other spectral methods that use other non-negative basis functions and their dual functions instead of Bernstein polynomials.

References

1. Varga, R.: On a discrete maximum principle. J. SIAM Numer. Anal. **3**, 355–359 (1966)
2. Ciarlet, P.G.: Discrete maximum principle for finite-difference operators. Aequ. Math. **4**, 266–268 (1970)
3. Drăgănescu, A., Dupont, T., Scott, L.R.: Failure of the discrete maximum principle for an elliptic finite element problem. Math. Comput. **74**(249), 1–23 (2005)
4. Jani, M., Javadi, S., Babolian, E., Bhatta, D.: Bernstein dual-Petrov-Galerkin method: application to 2D time fractional diffusion equation. Comput. Appl. Math. **37**, 2335–2353 (2018)
5. Foupouagnigni, M., Wouodjié, M.M.: On multivariate Bernstein polynomials. Mathematics **8**, 1397 (2020)
6. Hamadneh, T., Merker, J., Schimmel, W., Schuldt, G.: Simplicial Bernstein form and positivity certificates for solutions obtained in a stationary digital twin by Bernstein Bubnov-Galerkin method. In: Proceedings of ICoMS 2022, pp. 41–46. ACM, New York, USA (2022)
7. Gilbarg, D., Trudinger, N.: Elliptic Partial Differential Equations of Second Order. Springer (2001)
8. Narkawicz, A., Garloff, J., Smith, A.P., Munoz, C.A.: Bounding the range of a rational function over a box. Reliab. Comput. **17**, 34–39 (2012)

9. Garloff, J.: Convergent bounds for the range of multivariate polynomials. In: Proceedings of the International Symposium on Interval Mathematics 1985, LNCS, vol. 212, pp. 37–56. Springer (1986)
10. Ciesielski, Z.: The basis of B-splines in the space of algebraic polynomials. Ukr. Math. J. **38**, 311–315 (1986)
11. Jüttler, B.: The dual basis functions for the Bernstein polynomials. Adv. Comput. Math. **8**, 345–352 (1998)
12. Jani, M., Babolian, E., Javadi, S., Bhatta, D.: Banded operational matrices for Bernstein polynomials and application to the fractional advection-dispersion equation. Numer. Algorithms (2017)
13. Plemmons, R.J.: M-matrix characterizations. I - nonsingular M-matrices. Linear Algebr. Appl. **18**, 175–188 (1977)
14. Varga, R.: Matrix Iterative Analysis. Prentice Hall (1962)