

Study on Matrix Analysis for Steady Problems

Priyanka¹ Dr. Vakul Bansal²

¹Research Scholar, CMJ University, Shillong, Meghalaya

²Professor

Abstract - If the solution of a given boundary value problem satisfies a maximum principle, then a properly designed approximation should behave in the same way. A numerical scheme that does not generate spurious global extrema in the interior of the computational domain is said to satisfy a discrete maximum principle (DMP). As in the continuous case, the precise formulation of this criterion is problem-dependent. In particular, the zero row sum property (second rule from Section 1.6.3) has the same implications as the constraint $\sum v = 0$ in continuous maximum principles.

In the context of finite difference approximations to linear elliptic problems, sufficient conditions of DMP were formulated and proven by Varga [340] as early as in 1966. These conditions are related to the concept of monotone operators and, in particular, M-matrices which play an important role in numerical linear algebra [339, 354]. A general approach to DMP analysis for finite difference operators was developed by Ciarlet [63]. Its extension to finite elements in [64] features a proof of uniform convergence, as well as simple geometric conditions that ensure the validity of DMP for a piecewise-linear Galerkin discretization of the (linear) model problem.

Key Words: Computational, Monotone Operators, Uniform Convergence.

INTRODUCTION

On a triangular mesh under the assumption that $r \geq 0$ in Ω . The results obtained in [63, 64, 339] have illustrated the significance of DMPs for the analysis and design of numerical approximations. Various extensions and generalizations were published during the past three decades, see [61, 100, 151, 179, 180, 341, 311] and references therein. The frequently cited monograph by Ikeda [167] is devoted entirely to DMP for finite element models of convection-diffusion phenomena. Some low-order approximations of transport equations are known to satisfy a DMP unconditionally or under rather mild restrictions on the angles or aspect ratios of mesh cells. However, most *a priori* proofs are based on a set of sufficient conditions which become overly restrictive in the case of higher-order discretizations, singularly perturbed convection-diffusion equations, and anisotropic diffusion problems. A possible remedy to this problem is proposed in the next chapter. In this section, we review the algebraic constraints that ensure DMP and/or positivity preservation for steady transport problems of elliptic and hyperbolic type. A brief summary of the corresponding geometric conditions will also be presented.

REVIEW OF LITERATURE

A key ingredient of the mathematical theory behind the discrete maximum principles and positivity preservation is the following monotonicity concept [63, 339].

Definition 3.8. A regular matrix $A = \{a_{ij}\}$ is called *monotone* if $A^{-1} \geq 0$.

This kind of monotonicity is equivalent to the requirement that, for any vector u , $Au \geq 0 \Rightarrow u \geq 0$.

Matrix Analysis for Steady Problems 109 Of course, it is impractical to compute the inverse of A and check the sign of its entries. Instead, we will deal with a special class of matrices which are known to be monotone under certain constraints on the sign and magnitude of their coefficients.

Definition 3.9. A regular matrix $A = \{a_{ij}\}$ is called an *M-matrix* if $A^{-1} \geq 0$ and

$$a_{ij} \leq 0, \quad 8 \ j \neq i.$$

In other words, an M -matrix is a monotone matrix with nonpositive off-diagonal entries. These properties ensure positivity and convergence of iterative solvers.

Definition 3.10. A matrix $A = \{a_{ij}\}$ is called *diagonally dominant* (by rows) if $|a_{ii}| \geq \sum_{j \neq i} |a_{ij}|$, $8i$. (3.53)

Such a matrix is called *strictly diagonally dominant* if all inequalities are strict $|a_{ii}| > \sum_{j \neq i} |a_{ij}|$, $8i$. (3.54)

Definition 3.11. A matrix $A = \{a_{ij}\}$ of size $N \times N$ is called *irreducible* if there is no $N \times N$ permutation matrix P such that the following transformation is possible $PAPT = \begin{pmatrix} A_{11} & A_{12} & 0 \\ 0 & A_{22} & \end{pmatrix}$, where the size of A_{11} is $M \times M$, the size of A_{22} is $(N-M) \times (N-M)$, and $1 \leq M < N$.

It turns out that a matrix A is irreducible if and only if its directed graph is strongly connected ([339], p. 20) or, equivalently, if and only if for any i and $j \neq i$ there is a sequence of distinct indices $i = n_0, n_1, \dots, n_l = j$ such that $a_{n_{k-1}n_k} > 0$, $1 \leq k \leq l$.

MATERIAL AND METHOD

Consider the steady transport-reaction equation (3.9) discretized by a finite difference, finite volume, or finite element scheme. Let the approximate solution u_h , where the subscript h refers to the mesh size, be determined by a finite number $\bar{N}$ of degrees of freedom $u_1, \dots, u_{\bar{N}}$ that represent pointwise nodal values, control volume averages, or coefficients of piecewise-polynomial basis functions, respectively.

Hence, all information about the solution u_h can be packed into a vector $u \in \mathbb{R}^{\bar{N}}$.

Furthermore, the differential operator L acting on functions defined at infinitely many locations is replaced by a discrete operator L_h acting on vectors of length $\bar{N}$

$L_h : \mathbb{R}^{\bar{N}} \rightarrow \mathbb{R}^{\bar{N}}$

Regardless of the underlying approximation technique, we define this mapping as $L_h u = Au$, where $A = \{a_{ij}\}$ is a sparse $\bar{N} \times \bar{N}$ matrix and $u = \{u_i\}$ is the vector of nodal values.

The sparsity pattern of A depends on the mesh, on the type of discretization, and on the numbering of nodes. Since some nodal values are known from the Dirichlet boundary conditions, the size of the algebraic system reduces accordingly.

Let the first N nodes be associated with the unknown degrees of freedom, and the rest with the Dirichlet boundary values. This numbering convention implies that the 108 3 Maximum Principles

The subscripts i and j refer to row/column numbers from $N_i = \{1, \dots, N\}$ and $N_j = \{N+1, \dots, \bar{N}\}$, respectively. Thus, $u_i = \{u_1, \dots, u_N\}$ is the vector of unknowns, whereas $u_j = \{u_{N+1}, \dots, u_{\bar{N}}\}$ is given by the prescribed boundary data $u_j = g$. (3.49)

In this notation, the system of algebraic equations for the components of u_i reads $A_{ii}u_i = b_i - A_{ij}u_j$, (3.50) where b_i is the contribution of sources and Neumann boundary conditions, if any.

In a practical implementation, it is convenient to incorporate the Dirichlet boundary conditions into the $\bar{N} \times \bar{N}$ matrix A and solve the extended linear system [63]

In other words, the $\bar{N} \times \bar{N}$ matrix $\bar{A}$ is constructed from A by setting $A_{ii} = 0$ and $A_{ij} = I$, where I denotes the identity matrix with $\bar{N} - N$ rows and columns.

If the solution of the continuous problem satisfies Theorem 3.5 or 3.7, it is natural to require that the maxima of u_i be bounded by those of $u_j = g$. Likewise, all nodal values should be nonnegative if Theorem 3.6 or 3.8 is applicable. To verify the validity of DMP, one needs to analyze the properties of the discrete operator $\bar{A}$. *Remark 3.8.* In the context of linear systems, irreducibility ensures that it is impossible to extract a subsystem that can be solved independently. Matrices that result from discretization of partial differential equations are irreducible in most cases.

Definition 3.12. A matrix $A = \{a_{ij}\}$ is *irreducibly diagonally dominant* if it is irreducible and diagonally dominant, with strict dominance for at least one row i .

The following theorem yields a set of sufficient conditions which are commonly employed in DMP analysis based on the M -matrix property of discrete operators.

Theorem 3.13. If $A = \{a_{ij}\}$ is a strictly or irreducibly diagonally dominant $N \times N$ matrix with $a_{ii} > 0$, $8i = 1, \dots, N$ and $a_{ij} \leq 0$, $8j \neq i$, then $A^{-1} \geq 0$.

110 3 Maximum Principles

Proof. A common approach to the proof of this theorem is based on the splitting $A = D - C$, where the diagonal part $D = \text{diag}(A) > 0$ is nonsingular and $C \geq 0$. The diagonal dominance makes it possible to prove that the spectral radius ρ of $B = D^{-1}C \geq 0$ satisfies $\rho(B) < 1$. This condition holds if and only if the series $(I - B)^{-1} = I + B + B^2 + B^3 + \dots$ converges, see [131, 339] for technical details. Hence, $A^{-1} = (I - B)^{-1}D^{-1} \geq 0$.

If all diagonal entries of A are strictly positive and there are no positive offdiagonal ones, then diagonal dominance (3.53) requires that all row sums be nonnegative. The following definition summarizes the corresponding sign conditions.

Definition 3.13. A matrix $A = \{a_{ij}\}$ is said to be of *nonnegative type* [64, 69] if $a_{ii} > 0$, $a_{ij} \leq 0$, $i, j = 1, \dots, N$, (3.55)

$$a_{ij} \leq 0, \quad i, j = 1, \dots, N, \quad (3.56)$$

$\square_j$

$$a_{ij} \leq 0, \quad i, j = 1, \dots, N, \quad (3.57)$$

Corollary 3.12. By Theorem 3.13, a nonnegative-type matrix A is an M -matrix if inequality (3.57) is strict or A is irreducible and (3.57) is strict for at least one row.

Note that conditions (3.55)–(3.56) impose the same constraints as the third basic rule from Section 1.6.3. The second basic rule is satisfied if (3.57) holds as equality.

Under the assumptions of Corollary 3.12, the nonnegativity conditions are sufficient (but not *necessary*) for the matrix A to be monotone. Some other useful criteria related to M -matrices and monotonicity can be found in [46, 132, 153, 311, 346].

DISCRETE MAXIMUM PRINCIPLES

Given a discretization of the form (3.51), the monotonicity of the matrix $\bar{A}$ makes it possible to prove discrete counterparts of all maximum, minimum, and comparison principles established in Section 3.1. The uniqueness of the solution vector u follows from the regularity of $\bar{A}$. The usual approach to the proof of monotonicity is based on Theorem 3.13 and Corollary 3.12 since the nonnegativity conditions (3.55)–(3.57) are easy to verify for an arbitrary space discretization of the transport equation. To prove that the solution of problem (3.51) attains its maximum on the set Γ_D of Dirichlet boundary nodes, we need a discrete counterpart of the incompressibility constraint $\nabla \cdot \mathbf{v} = 0$. At the continuous level, it implies that $Lu = L(u+c)$ for an arbitrary constant c . According to the second basic rule from Section 1.6.3, the

discrete operator A should have zero row sums to inherit this property. Thus, the *global* discrete maximum principle for nodal values can be formulated as follows.

CONCLUSION

That is, if the source term is absent and the solution has a constant value u_i at all neighboring nodes, then u_i must also assume this value. This property is guaranteed by the zero row sum condition (3.58). Only a poor discretization of the transport equation with $\nabla \cdot \mathbf{v} = 0$ would produce $u_i \neq u$ in this situation ([268], p. 39).

Since estimate (3.63) holds for any interior node, successive application of the local DMP can be used to prove (3.59) if A is irreducible, cf. [69, 344]. If the row sums of $\bar{A}$ are nonvanishing for some $i \in \Gamma_D$, then the DMP may cease to hold but positivity preservation can be inferred from the fact that $\bar{A}$ is monotone. If the matrix $\bar{A}$ is given by (3.52), where A is monotone and $A \geq 0$, then discretization (3.51) is positivity-preserving, that is, $b_i \geq 0$, $u_i \geq 0$. (3.64) *Proof.* The matrix $\bar{A}$ given by (3.52) is regular if and only if the block A is regular. Furthermore, the inverse matrices. Some finite element approximations to elliptic problems like (3.47) are known to satisfy the DMP conditions (3.55)–(3.57) on a suitably designed mesh. The derivation of geometric constraints that ensure monotonicity has been one of the primary research directions in the DMP analysis for finite element schemes [64, 100, 179,

180]. Below, we present some useful geometric criteria in the context of linear and bilinear Galerkin discretizations of the Laplace operator in two space dimensions.

Definition 3.14. A triangular mesh is called *strongly acute* if all angles are smaller than $\pi/2$ and *weakly acute* (or *nonobtuse*) if all angles are not greater than $\pi/2$.

Theorem 3.21. The discrete Laplace operator $\bar{A}$ for the linear finite element approximation on a triangular mesh of weakly acute type is monotone [18, 64]. This classical result dates back to the paper by Ciarlet and Raviart [64]. In the 3D case, a tetrahedral mesh is said to be of acute type if all internal angles between the faces of tetrahedra are not greater than $\pi/2$. Again, this condition ensures that the discrete Laplace operator is monotone if linear finite elements are employed [189].

Definition 3.15. A triangular mesh is a *Delaunay triangulation* if no vertex of this mesh is inside the circumcircle of any triangle to which it does not belong.

Theorem 3.22. *The discrete Laplace operator $\bar{A}$ for the linear finite element approximation on a Delaunay triangulation is monotone [18].*

Delaunay triangulations maximize the minimum angle as to avoid excessively stretched triangles. It is known that there exists a unique Delaunay triangulation for 114 3 Maximum Principles any set of points that do not lie on the same line. Moreover, fast algorithms are available for creating such triangulations [55, 115, 226], which makes them very popular with finite element practitioners. Figure 3.4 displays a simple 2D Delaunay triangulation generated using the MATLAB function delaunay. In three dimensions, no vertex of a tetrahedron is inside the circumsphere of any other tetrahedron. However, the discrete Laplacian operator for a linear FEM approximation on an arbitrary 3D Delaunay triangulation may fail to be a matrix of nonnegative type [18]. This does not necessarily cause a violation of the DMP but it cannot be ruled out anymore.

Definition 3.16. A rectangular mesh is called *nonnarrow* if the ratio of longest and shortest mesh edge is not greater than p_2 for any rectangle [100].

Theorem 3.23. *The discrete Laplace operator $\bar{A}$ for the bilinear finite element approximation on a rectangular mesh of nonnarrow type is monotone [61, 100].*

This theorem explains why iterative solution techniques that rely on the M-matrix property may experience convergence problems when applied to discretizations of second-order PDEs on quadrilateral/hexahedral meshes with high aspect ratios. Geometric DMP conditions for various elliptic and parabolic problems have been formulated building on the above results [100, 116, 179, 180]. Even if convective effects are present, it is desirable to use a sufficiently regular mesh that satisfies the above conditions, so that at least the discrete diffusion operator poses no hazard to DMP. Moreover, it may offset a nonmonotone convective part if the Peclet number is small or a sufficiently large amount of artificial diffusion is added [51, 69]. Alternatively, the *upwind triangle* method or other techniques can be used to construct a monotone approximation of convective terms [11, 167, 200, 288, 348].

REFERENCES

1. M. Ainsworth and J.T. Oden, *A Posteriori Error Estimation in Finite Element Analysis*. John Wiley & Sons, New York, 2000.
2. M. Ainsworth, J.Z. Zhu, A.W. Craig, O.C. Zienkiewicz, Analysis of the Zienkiewicz-Zhu a posteriori error estimator in the finite element method. *Int. J. Numer. Methods Engrg.* **28:9** (1989) 2161–2174.
3. J.D. Anderson, Jr., *Computational Fluid Dynamics. The Basics with Applications*. McGraw- Hill, 1995.
4. J.D. Anderson, Jr., *Modern Compressible Flow: With Historical Perspective*, McGraw-Hill, 1990.
5. F. Angrand, A. Derieux, L. Loth, G. Vijayasundaram, Simulation of Euler transonic flows by means of explicit finite element type schemes. *INRIA Research Report* **250**, 1983.
6. F. Angrand and A. Derieux, Some explicit triangular finite element schemes for the Euler equations. *Int. J. Numer. Methods Fluids* **4** (1984) 749–764.
7. P. Arminjon and A. Dervieux, Construction of TVD-like artificial viscosities on 2-dimensional arbitrary FEM grids. *INRIA Research Report* **1111**, 1989.
8. M. Arora and P.L. Roe, A well-behaved TVD limiter for high-resolution calculations of unsteady flow. *J. Comput. Phys.* **132** (1997) 3–11.
9. Athena test suite, <http://www.astro.virginia.edu/VITA/ATHENA/dmr.html>.
10. K. Baba and M. Tabata, On a conservative upwind finite element scheme for convective diffusion equations. *RAIRO Numerical Analysis* **15** (1981) 3–25.
11. I. Babuška and W.C. Rheinboldt, Error estimates for adaptive finite element computations. *SIAM J. Numer. Anal.* **15:4** (1978) 736–354.
12. I. Babuška and W.C. Rheinboldt, A posteriori error estimates for the finite element method. *Int. J. Numer. Methods Engrg.* **12:10** (1978) 1597–1615.
13. W. Bangerth and R. Rannacher, Adaptive finite element methods for differential equations. Lectures in Mathematics, ETH Zürich, Birkhäuser, 2003.
14. R.E. Bank, A.H. Sherman, A. Weiser, Some refinement algorithms and data structures for regular local mesh refinement. In: R. Stepleman (ed.), *Scientific Computing, Applications of Mathematics and Computing to the Physical Sciences*, IMACS Transactions on Scientific

Computation, Vol. I. North-Holland, Amsterdam, 1983, 3–17.

15. R.E. Bank and R.K. Smith, Mesh smoothing using a posteriori error estimates. *SIAM J. Numer. Anal.* **34** (1997) 979–997.
16. T.J. Barth, Numerical aspects of computing viscous high Reynolds number flows on unstructured meshes. *AIAA Paper*, 91-0721, 1991.
17. T.J. Barth, Aspects of unstructured grids and finite volume solvers for the Euler and Navier- Stokes equations. In: Lecture Series 1994-05, von Karman Institute for Fluid Dynamics, Brussels, 1994.