ARTICLE IN PRESS

Journal of Computational and Applied Mathematics xxx (xxxx) xxx

Contents lists available at ScienceDirect

Journal of Computational and Applied Mathematics

journal homepage: www.elsevier.com/locate/cam

Conditioning of implicit Runge–Kutta integration for finite element approximation of linear diffusion equations on anisotropic meshes☆

Weizhang Huang [a], Lennard Kamenski [b,*], Jens Lang [c]

[a] Department of Mathematics, The University of Kansas, Lawrence, KS 66045, USA
[b] m4sim GmbH, 10117 Berlin, Germany
[c] Department of Mathematics, TU Darmstadt, 64293 Darmstadt, Germany

ARTICLE INFO

Article history:
Received 2 February 2019
Received in revised form 14 August 2019

MSC:
65M60
65M50
65F35
65F15

Keywords:
Finite element method
Anisotropic mesh
Condition number
Parabolic equation
Implicit Runge–Kutta method

ABSTRACT

The conditioning of implicit Runge–Kutta (RK) integration for linear finite element approximation of diffusion equations on general anisotropic meshes is investigated. Bounds are established for the condition number of the resulting linear system with and without diagonal preconditioning for the implicit Euler (the simplest implicit RK method) and general implicit RK methods. Two solution strategies are considered for the linear system resulting from general implicit RK integration: the simultaneous solution where the system is solved as a whole and a successive solution which follows the commonly used implementation of implicit RK methods to first transform the system into a number of smaller systems using the Jordan normal form of the RK matrix and then solve them successively.

For the simultaneous solution in case of a positive semidefinite symmetric part of the RK coefficient matrix and for the successive solution it is shown that the conditioning of an implicit RK method behaves like that of the implicit Euler method. If the smallest eigenvalue of the symmetric part of the RK coefficient matrix is negative and the simultaneous solution strategy is used, an upper bound on the time step is given so that the system matrix is positive definite.

The obtained bounds for the condition number have explicit geometric interpretations and take the interplay between the diffusion matrix and the mesh geometry into full consideration. They show that there are three mesh-dependent factors that can affect the conditioning: the number of elements, the mesh nonuniformity measured in the Euclidean metric, and the mesh nonuniformity with respect to the inverse of the diffusion matrix. They also reveal that the preconditioning using the diagonal of the system matrix, the mass matrix, or the lumped mass matrix can effectively eliminate the effects of the mesh nonuniformity measured in the Euclidean metric. Illustrative numerical examples are given.

© 2019 Elsevier B.V. All rights reserved.

☆ Supported in part by the DFG, Germany under grant KA 3215/2-1 and the Darmstadt Graduate Schools of Excellence *Computational Engineering* and *Energy Science and Engineering, Germany*.

* Corresponding author.
E-mail addresses: whuang@ku.edu (W. Huang), lkamenski@m4sim.de (L. Kamenski), lang@mathematik.tu-darmstadt.de (J. Lang).

https://doi.org/10.1016/j.cam.2019.112497
0377-0427/© 2019 Elsevier B.V. All rights reserved.

Please cite this article as: W. Huang, L. Kamenski and J. Lang, Conditioning of implicit Runge–Kutta integration for finite element approximation of linear diffusion equations on anisotropic meshes, Journal of Computational and Applied Mathematics (2019) 112497, https://doi.org/10.1016/j.cam.2019.112497.

1. Introduction

The nonuniformity of adaptive meshes has considerable effects on the conditioning of the discrete approximation of partial differential equations (PDEs) and their efficient solution. To study these effects, we investigate the implicit Runge–Kutta (RK) integration for the linear finite element (FE) approximation of linear diffusion equations on general simplicial anisotropic meshes for the initial–boundary value problem (IBVP)

$$\begin{cases} \partial_t u = \nabla \cdot (\mathbb{D}\nabla u)\,, & \boldsymbol{x} \in \Omega, \quad t > 0, \\ u(\boldsymbol{x}, t) = 0, & \boldsymbol{x} \in \partial\Omega, \quad t > 0, \\ u(\boldsymbol{x}, 0) = u_0(\boldsymbol{x}), & \boldsymbol{x} \in \Omega, \end{cases} \tag{1}$$

where $\Omega \subset \mathbb{R}^d$ ($d \geq 1$) is a bounded polygonal or polyhedral domain, u_0 is a given initial solution, and $\mathbb{D}$ is the diffusion matrix. We assume that $\mathbb{D} = \mathbb{D}(\boldsymbol{x})$ is time independent and symmetric and uniformly positive definite on Ω, i.e.,

$$\exists\, d_{\min}, d_{\max} > 0: \quad d_{\min} I \leq \mathbb{D}(\boldsymbol{x}) \leq d_{\max} I, \quad \forall \boldsymbol{x} \in \Omega, \tag{2}$$

where the less-than-or-equal sign means that the difference between the two matrices is positive semidefinite. We consider the Dirichlet boundary condition in this work but the analysis is applicable to other boundary condition types without major modifications.

Much effort has been made in the past to understand the effects of mesh nonuniformity on the conditioning of FE approximations. For example, Fried [1] obtains a bound on the condition number of the stiffness matrix for the linear FE approximation of the Laplace operator for general meshes. For the Laplace operator on isotropic adaptive grids, Bank and Scott [2] show that the condition number of the diagonally scaled stiffness matrix is essentially the same as for a regular mesh. Ainsworth, McLean, and Tran [3] and Graham and McLean [4] extend this result to the boundary element equations for locally quasi-uniform meshes and provide a bound in terms of patch volumes and aspect ratios. Du et al. [5] obtain a bound on the condition number of the stiffness matrix for a general diffusion operator on anisotropic meshes which reveals the relation between the condition number and some mesh quality measures. For the FE approximations of parabolic problems, Zhu and Du [6,7] develop mesh dependent stability and condition number estimates for the explicit and implicit Euler methods. In the case of a lumped mass matrix (M_{lump}), Shewchuk [8] provides a bound on the largest eigenvalue of $M_{lump}^{-1}A$ in terms of the maximum eigenvalues of local element matrices, where A is the stiffness matrix. The results mentioned above allow anisotropic adaptive meshes but do not fully take into account the interplay between the mesh geometry and the diffusion matrix. (An exception is the bound by Shewchuk which includes the effects of the diffusion coefficients.) Moreover, the existing analysis employs mesh restrictions in form of mesh regularity assumptions, e.g., local mesh uniformity, or parameters in final estimates that are related to mesh regularity, such as the maximum ratio of volumes of neighbouring elements or the maximum number of elements in a patch.

The objective of this work is to develop estimates on the conditioning of the resulting linear system that take the interplay between the mesh geometry and the diffusion matrix into full consideration, have explicit geometric interpretation, and make no prior assumptions on the mesh regularity. This is a continuation of our previous effort to develop bounds for the condition number of the stiffness matrix for the linear FE equations of a general diffusion operator on arbitrary anisotropic meshes [9,10] and the largest permissible time steps for explicit RK schemes for both linear and high order FE approximations of the IBVP (1) [11,12]. In particular, these bounds show [10] that the condition number of the stiffness matrix depends on three factors: the factor depending on the number of mesh elements and corresponding to the condition number of the linear FE equations for the Laplace operator on a uniform mesh, the nonuniformity of the mesh viewed in the metric defined by the inverse diffusion matrix, $\mathbb{D}^{-1}$, and the mesh nonuniformity measured in the Euclidean metric. Moreover, the Jacobi preconditioning, an optimal diagonal scaling for a symmetric positive definite sparse matrix [13, Corollary 7.6ff.], can effectively eliminate the effects of mesh nonuniformity and reduce those of the mesh nonuniformity with respect to $\mathbb{D}^{-1}$ [10]. Detailed characterizations of the condition number according to the mesh concentration distribution can be obtained using Green's functions [1,9].

We first consider the implicit Euler method (the simplest implicit RK method) and establish bounds for the condition number of the corresponding system matrix with and without diagonal scaling.

For general implicit RK methods, we consider two strategies for solving the resulting system: the simultaneous solution and a successive solution. We need to point out that for general implicit RK methods, the coefficient matrix is not necessarily symmetric and thus a condition number in the standard definition will not provide much information for iterative solvers. Motivated by the convergence estimates by Eisenstat et al. [14] for the generalized minimal residual method (GMRES) for nonsymmetric systems, we use here the ratio of the maximum singular value of the coefficient matrix and the minimum eigenvalue of its symmetric part to measure the conditioning.

For the simultaneous solution, the system is solved as a whole. Obtained estimates reveal that the conditioning of implicit RK methods with positive semidefinite symmetric part of the coefficient matrix behaves like that of the implicit Euler method (Theorem 5.1 in Section 5). If the smallest eigenvalue of the symmetric part of the RK coefficient matrix is negative, we provide an upper bound (44) on the possible time step so that the system matrix is positive definite. This condition can be serious and lead to $\Delta t = \mathcal{O}(h^2)$.

For the successive solution, which follows the commonly used implementation of implicit RK methods [15,16], the system is first transformed into a number of smaller systems using the Jordan normal form of the RK matrix and then

Please cite this article as: W. Huang, L. Kamenski and J. Lang, Conditioning of implicit Runge–Kutta integration for finite element approximation of linear diffusion equations on anisotropic meshes, Journal of Computational and Applied Mathematics (2019) 112497, https://doi.org/10.1016/j.cam.2019.112497.

solved successively. We show that the conditioning of the implicit RK integration behaves like that of the implicit Euler method and is determined by the conditioning of two types of matrices. The first one is similar to the implicit Euler method and corresponds to the real eigenvalues of the RK matrix, while the second, twice as large, corresponds to the complex eigenvalues of the RK matrix (cf. (51) in Section 5).

The paper is organized as follows. We first introduce the FE formulation and its implicit RK integration (Section 2) and provide preliminary estimates for the extremal eigenvalues of the mass and stiffness matrices (Section 3). The main results for the conditioning of the coefficient matrices are given in Sections 4 and 5, followed by numerical examples (Section 6) and conclusions (Section 7).

2. Linear FE approximation and implicit RK integration

Let $\{\mathcal{T}_h\}$ be a given family of simplicial meshes for the domain Ω and N, N_v, and N_{vi} the number of mesh elements, vertices, and interior vertices, respectively. For convenience, we assume that the vertices are ordered such that the first N_{vi} vertices are the interior ones. The element patch associated with the jth vertex is denoted by ω_j, K denotes a given mesh element and $\hat{K}$ is the reference element which is assumed to have been taken as a unitary equilateral simplex. Element and patch volumes are denoted by $|K|$ and $|\omega_j| = \sum_{K\in\omega_j} |K|$. For each mesh element $K \in \mathcal{T}_h$, we denote the invertible affine mapping from $\hat{K}$ to K and its Jacobian matrix by F_K and F_K', respectively. Note that F_K' is a constant matrix and $\det(F_K') = |K|$.

Let $V^h \subset H_0^1(\Omega)$ be the linear FE space associated with $\mathcal{T}_h$. The piecewise linear FE solution $u_h(t) \in V^h$, $t > 0$ for (1) is defined by

$$\int_\Omega \frac{\partial u_h}{\partial t} v_h \, \mathrm{d}\boldsymbol{x} = -\int_\Omega (\nabla v_h)^T \mathbb{D} \nabla u_h \, \mathrm{d}\boldsymbol{x}, \qquad \forall v_h \in V^h, \quad t > 0, \tag{3}$$

subject to the initial condition

$$\int_\Omega u_h(\boldsymbol{x}, 0) v_h \, \mathrm{d}\boldsymbol{x} = \int_\Omega u^0(\boldsymbol{x}) v_h \, \mathrm{d}\boldsymbol{x}, \qquad \forall v_h \in V^h.$$

It can be expressed as

$$u_h(\boldsymbol{x}, t) = \sum_{j=1}^{N_{vi}} u_j(t) \phi_j(\boldsymbol{x}),$$

where ϕ_j is the linear basis function associated with the jth vertex. Inserting this into (3) and taking $v_h = \phi_k$, $k = 1, \ldots, N_{vi}$, successively yields the system

$$M \frac{d\boldsymbol{u}}{dt} = -A\boldsymbol{u}, \tag{4}$$

where $\boldsymbol{u} = (u_1, \ldots, u_{N_{vi}})^T$ and M and A are the mass and stiffness matrices with

$$M_{kj} = \int_\Omega \phi_k \phi_j \, \mathrm{d}\boldsymbol{x} \quad \text{and} \quad A_{kj} = \int_\Omega \nabla\phi_k \cdot \mathbb{D} \nabla \phi_j \, \mathrm{d}\boldsymbol{x}, \qquad k, j = 1, \ldots, N_{vi}. \tag{5}$$

For the time integration of (4) we consider a general implicit s-stage RK method with the Butcher tableau

c_1	γ_{11}	γ_{12}	$\cdots$	γ_{1s}
c_2	γ_{21}	γ_{22}	$\cdots$	γ_{2s}
$\vdots$	$\vdots$	$\vdots$		$\vdots$
c_s	γ_{s1}	γ_{s2}	$\cdots$	γ_{ss}
	b_1	b_2	$\cdots$	b_s

and assume that the eigenvalues of the RK matrix have nonnegative real parts. This requirement is satisfied by most implicit RK methods (e.g., see Table 1). Applying the method to (4) yields

$$M\boldsymbol{v}_k + \Delta t A \sum_{j=1}^{s} \gamma_{kj} \boldsymbol{v}_j = -A\boldsymbol{u}^n, \quad k = 1, \ldots, s, \tag{6}$$

$$\boldsymbol{u}^{n+1} = \boldsymbol{u}^n + \Delta t \sum_{k=1}^{s} b_k \boldsymbol{v}_k, \tag{7}$$

where $\boldsymbol{u}^n$ and $\boldsymbol{u}^{n+1}$ are the approximations of $\boldsymbol{u}(t_n)$ and $\boldsymbol{u}(t_{n+1})$. The major cost of finding $\boldsymbol{u}^{n+1}$ in the above method is the solution of (6) for $\boldsymbol{v}_1, \ldots, \boldsymbol{v}_s$. Using the Kronecker matrix product $\otimes$ (e.g., see [17]), the $s \times s$ identity matrix I_s, and

Please cite this article as: W. Huang, L. Kamenski and J. Lang, Conditioning of implicit Runge–Kutta integration for finite element approximation of linear diffusion equations on anisotropic meshes, Journal of Computational and Applied Mathematics (2019) 112497, https://doi.org/10.1016/j.cam.2019.112497.

Table 1
Eigenvalues and singular values of the implicit RK coefficient matrix Γ and the minimum eigenvalue of $(\Gamma+\Gamma^T)/2$.

Method	Order	Eigenvalues	$\sigma_{\max}$	$\sigma_{\min}$	$\lambda_{\min}((\Gamma+\Gamma^T)/2)$
Gauss	4	$0.25 \pm 0.1443i$	0.6319	0.1319	0
Gauss	6	$0.2153,\ 0.1423 \pm 0.1358i$	0.6629	0.0635	−0.0563
Radau IA	3	$0.3333 \pm 0.2357i$	0.5	0.3333	0.25
Radau IIA	5	$0.2749,\ 0.1626 \pm 0.1849i$	0.8023	0.0923	−0.0822
Lobatto IIIA	4	$0,\ 0.25 \pm 0.1443i$	0.7947	0	−0.0736

$\Gamma := (\gamma_{kj})_{k,j=1}^{s}$, the coefficient matrix of (6) can be expressed as

$$\mathbb{A} \equiv I_s \otimes M + \Delta t\, \Gamma \otimes A. \tag{8}$$

The simplest implicit RK method is the implicit Euler method with $s = 1$, $\Gamma = 1$, $c_1 = 1$, and $b_1 = 1$ for which (6) and (7) are reduced to

$$(M + \Delta t\, A)\boldsymbol{u}^{n+1} = M\boldsymbol{u}^n. \tag{9}$$

In Section 4 we study the conditioning of the coefficient matrix $M + \Delta t\, A$ related to the efficient iterative solution of (9). The system (9) can also be solved using a symmetric and positive definite preconditioner P, which leads to

$$P^{-\frac{1}{2}}(M + \Delta t\, A)P^{-\frac{1}{2}}\boldsymbol{w}^{n+1} = P^{-\frac{1}{2}}MP^{-\frac{1}{2}}\boldsymbol{w}^n, \qquad \boldsymbol{w}^n = P^{\frac{1}{2}}\boldsymbol{u}^n. \tag{10}$$

The efficient iterative solution of (10) is related to the conditioning of $P^{-\frac{1}{2}}(M + \Delta t\, A)P^{-\frac{1}{2}}$. In this work, we consider as preconditioners the mass matrix M, its diagonal part M_D, the lumped mass matrix M_{lump}, and the diagonal part $M_D + \Delta t\, A_D$ of $M + \Delta t\, A$. The choice $P = M$ is of theoretical importance but impractical since $M^{-\frac{1}{2}}$ is expensive to compute. The other choices are simple and economic to implement.

The results for the Euler method will be used in Section 5 to study the conditioning of (8) and its diagonally preconditioned version for general implicit RK methods.

3. Preliminary estimates on the extreme eigenvalues of the mass and stiffness matrices

Hereafter, C denotes a generic constant which may have different values at different appearances and may depend on the dimension, the choice of the reference element, and the reference basis linear functions but is independent of the mesh and the IBVP coefficients. For notation simplicity, when using this generic constant C, we will sometime write $a \gtrsim b$ and $a \lesssim b$ meaning $a \geq Cb$ and $a \leq Cb$, respectively.

Lemma 3.1 ([10, Proof of Theorem 3.1]). *The mass matrix M and its diagonal part M_D are related by*

$$\frac{1}{2}M_D \leq M \leq \left(1 + \frac{d}{2}\right)M_D \quad \text{and} \quad M_{jj} = \frac{2|\omega_j|}{(d+1)(d+2)}. \tag{11}$$

Lemma 3.2 ([12, Lemma 2.3]). *The mass matrix M and the lumped mass matrix M_{lump} are related by*

$$\frac{1}{d+2}M_{lump} \leq M \leq \frac{d+2}{2}M_{lump}. \tag{12}$$

Lemma 3.3 ([10, Lemma 4.1]). *The stiffness matrix A and its diagonal part A_D satisfy*

$$A \leq (d+1)A_D. \tag{13}$$

Lemma 3.4 ([12, Lemma 2.5]). *The diagonal entries of the stiffness matrix are bounded by*

$$C_{\hat{\nabla}} \sum_{K\in\omega_j} |K| \lambda_{\min}\left((F_K')^{-1}\mathbb{D}_K(F_K')^{-T}\right) \leq A_{jj} \leq C_{\hat{\nabla}} \sum_{K\in\omega_j} |K| \lambda_{\max}\left((F_K')^{-1}\mathbb{D}_K(F_K')^{-T}\right), \tag{14}$$

where $\mathbb{D}_K = \frac{1}{|K|}\int_K \mathbb{D}(\boldsymbol{x})\, d\boldsymbol{x}$ is the average of $\mathbb{D}$ on K and $C_{\hat{\nabla}} = \frac{d+1}{d}\left(\frac{d!}{\sqrt{d+1}}\right)^{\frac{2}{d}}$.

The next two lemmas establish bounds for A_{jj} with a more explicit geometric interpretation than (14). For this, we denote the diameter and the minimal height of K in the metric $\mathbb{D}_K^{-1}$ by $h_{K,\mathbb{D}^{-1}}$ and $a_{K,\mathbb{D}^{-1}}$, respectively. The average element diameter is defined as

$$h_{\mathbb{D}^{-1}} = \left(\frac{|\Omega|_{\mathbb{D}^{-1}}}{N}\right)^{\frac{1}{d}}, \quad |\Omega|_{\mathbb{D}^{-1}} = \sum_{K\in\mathcal{T}_h} |K|\sqrt{\det(\mathbb{D}_K^{-1})}. \tag{15}$$

Please cite this article as: W. Huang, L. Kamenski and J. Lang, Conditioning of implicit Runge–Kutta integration for finite element approximation of linear diffusion equations on anisotropic meshes, Journal of Computational and Applied Mathematics (2019) 112497, https://doi.org/10.1016/j.cam.2019.112497.

Further, let $\hat{h}$, $\hat{\rho}$, and $\hat{a}$ be the diameter, the in-diameter, and the minimal height of the unitary equilateral $\hat{K}$, respectively, i.e.,

$$\hat{h} = \sqrt{2}\left(\frac{d!}{\sqrt{d+1}}\right)^{\frac{1}{d}}, \quad \hat{\rho} = \sqrt{\frac{2}{d(d+1)}}\hat{h}, \quad \hat{a} = \sqrt{\frac{d+1}{2d}}\hat{h}.$$

Lemma 3.5 (*[18, Lemmas 4.1 and 4.2]*)**.** *It holds*

$$\frac{h_{K,\mathbb{D}^{-1}}^2}{\hat{h}^2} \le \|(F_K')^T \mathbb{D}_K^{-1} F_K'\|_2 \le \frac{h_{K,\mathbb{D}^{-1}}^2}{\hat{\rho}^2}, \tag{16}$$

$$\frac{\hat{a}^2}{a_{K,\mathbb{D}^{-1}}^2} \le \|(F_K')^{-1} \mathbb{D}_K (F_K')^{-T}\|_2 \le \frac{d^2\hat{a}^2}{a_{K,\mathbb{D}^{-1}}^2}. \tag{17}$$

Lemma 3.6. *It holds*

$$\hat{\rho}^2 C_{\hat{\mathbb{V}}} h_{\mathbb{D}^{-1}}^{-2} \sum_{K\in\omega_j} |K| \left(\frac{h_{\mathbb{D}^{-1}}}{h_{K,\mathbb{D}^{-1}}}\right)^2 \le A_{jj} \le d^2\hat{a}^2 C_{\hat{\mathbb{V}}} h_{\mathbb{D}^{-1}}^{-2} \sum_{K\in\omega_j} |K| \left(\frac{h_{\mathbb{D}^{-1}}}{a_{K,\mathbb{D}^{-1}}}\right)^2. \tag{18}$$

Proof. Combining (14) and (17) yields

$$A_{jj} \le d^2\hat{a}^2 C_{\hat{\mathbb{V}}} \sum_{K\in\omega_j} |K| a_{K,\mathbb{D}^{-1}}^{-2} = d^2\hat{a}^2 C_{\hat{\mathbb{V}}} h_{\mathbb{D}^{-1}}^{-2} \sum_{K\in\omega_j} |K| \left(\frac{h_{\mathbb{D}^{-1}}}{a_{K,\mathbb{D}^{-1}}}\right)^2,$$

which gives the right inequality of (18). On the other hand, (16) gives

$$\lambda_{\min}((F_K')^{-1}\mathbb{D}_K(F_K')^{-T}) = \|(F_K')^T \mathbb{D}_K^{-1} F_K'\|_2^{-1} \ge \hat{\rho}^2 h_{K,\mathbb{D}^{-1}}^{-2}.$$

Combining this with (14) leads to the left inequality of (18). □

A mesh that is uniform in the metric $\mathbb{D}^{-1}$ (*a $\mathbb{D}^{-1}$-uniform mesh*) satisfies $h_{K,\mathbb{D}^{-1}} \sim a_{K,\mathbb{D}^{-1}} \sim h_{\mathbb{D}^{-1}}$. In such a case, (18) implies $A_{jj} \sim |\omega_j| h_{\mathbb{D}^{-1}}^{-2} \sim |\omega_j| N^{\frac{2}{d}}$.

Lemma 3.7 (*[10, Lemma 5.1]*)**.** *The smallest eigenvalue of the stiffness matrix is bounded by*

$$\lambda_{\min}(A) \gtrsim d_{\min} N^{-1} \Psi_E^{-1}, \tag{19}$$

where

$$\Psi_E = \begin{cases} 1, & d = 1, \\ 1 + \ln\left(\frac{|\bar{K}|}{|K_{\min}|}\right), & d = 2, \\ \left(\frac{1}{N}\sum_{K\in\mathcal{T}_h} \left(\frac{|\bar{K}|}{|K|}\right)^{\frac{d-2}{2}}\right)^{\frac{2}{d}}, & d \ge 3, \end{cases} \tag{20}$$

$|K_{\min}|$ and $|\bar{K}| = \frac{1}{N}|\Omega|$ are the minimal and the average element volumes and $d_{\min}$ is the global smallest eigenvalue of $\mathbb{D}$ (cf. (2)).

The factor Ψ_E shows that the dependence of the lower bound on the mesh non-uniformity is mild although getting stronger in higher dimensions: Ψ_E is mesh-independent in 1d, contains only $\ln\left(|\bar{K}|/|K_{\min}|\right)$ in 2d, and involves the generalized $d/2$-mean of the ratio $\left(|\bar{K}|/|K|\right)^{(d-2)/d}$ over all elements for $d \ge 3$.

The next lemmas summarize bounds for the extremal eigenvalues of M and A.

Lemma 3.8. *The extremal eigenvalues of $P^{-\frac{1}{2}} M P^{-\frac{1}{2}}$ are bounded by*

$$\lambda_{\max}(P^{-\frac{1}{2}} M P^{-\frac{1}{2}}) \le \begin{cases} \frac{|\omega_{\max}|}{d+1}, & P = I_{N_{vi}}, \\ 1 + \frac{d}{2}, & P = M_D,\ M_{lump},\ M_D + \Delta t\, A_D, \end{cases} \tag{21}$$

and

$$\lambda_{\min}(P^{-\frac{1}{2}}MP^{-\frac{1}{2}}) \geq \begin{cases} \frac{|\omega_{\min}|}{(d+1)(d+2)}, & P = I_{N_{vi}}, \\ 1/2, & P = M_D, \\ 1/(d+2), & P = M_{lump}, \\ C\left(1 + \Delta t\, h_{\mathbb{D}^{-1}}^{-2} \max_j \sum_{K\in\omega_j} \frac{|K|}{|\omega_j|}\left(\frac{h_{\mathbb{D}^{-1}}}{a_{K,\mathbb{D}^{-1}}}\right)^2\right)^{-1} & P = M_D + \Delta t\, A_D, \end{cases} \tag{22}$$

where $|\omega_{\max}|$ and $|\omega_{\min}|$ are the maximal and minimal patch volumes, respectively.

Proof. The inequalities (21) and (22) for $P = I_{N_{vi}}$ (no preconditioning), $P = M_D$, and M_{lump} follow from Lemmas 3.1 and 3.2. For $P = M_D + \Delta t\, A_D$, (11) and (18) give

$$P_{jj} \lesssim \left|\omega_j\right| + \Delta t\, h_{\mathbb{D}^{-1}}^{-2} \sum_{K\in\omega_j} |K| \left(\frac{h_{\mathbb{D}^{-1}}}{a_{K,\mathbb{D}^{-1}}}\right)^2. \tag{23}$$

Then, (22) follows from

$$\boldsymbol{v}^T P^{-\frac{1}{2}} M P^{-\frac{1}{2}} \boldsymbol{v} \gtrsim \sum_j \frac{v_j^2 \left|\omega_j\right|}{P_{jj}} \gtrsim \sum_j v_j^2 \left(1 + \Delta t\, h_{\mathbb{D}^{-1}}^{-2} \sum_{K\in\omega_j} \frac{|K|}{|\omega_j|}\left(\frac{h_{\mathbb{D}^{-1}}}{a_{K,\mathbb{D}^{-1}}}\right)^2\right)^{-1}. \quad \square$$

Lemma 3.9. *The extremal eigenvalues of $P^{-\frac{1}{2}}AP^{-\frac{1}{2}}$ are bounded by*

$$\lambda_{\max}(P^{-\frac{1}{2}}AP^{-\frac{1}{2}}) \leq \begin{cases} C h_{\mathbb{D}^{-1}}^{-2} \max_j \sum_{K\in\omega_j} |K| \left(\frac{h_{\mathbb{D}^{-1}}}{a_{K,\mathbb{D}^{-1}}}\right)^2, & P = I_{N_{vi}}, \\ C h_{\mathbb{D}^{-1}}^{-2} \max_j \sum_{K\in\omega_j} \frac{|K|}{|\omega_j|}\left(\frac{h_{\mathbb{D}^{-1}}}{a_{K,\mathbb{D}^{-1}}}\right)^2, & P = M,\ M_D,\ M_{lump}, \\ (d+1)\Delta t^{-1}, & P = M_D + \Delta t\, A_D, \end{cases} \tag{24}$$

and

$$\lambda_{\min}(P^{-\frac{1}{2}}AP^{-\frac{1}{2}}) \gtrsim \begin{cases} d_{\min} N^{-1} \Psi_E^{-1}, & P = I_{N_{vi}}, \\ \lambda_{\mathbb{D}}, & P = M,\ M_D,\ M_{lump}, \\ d_{\min} \Psi_{\mathbb{D}}^{-1}, & P = M_D + \Delta t\, A_D, \end{cases} \tag{25}$$

where $\lambda_{\mathbb{D}}$ is the minimal eigenvalue of the operator $-\nabla\cdot(\mathbb{D}\nabla)$ and

$$\Psi_{\mathbb{D}} = \begin{cases} 1 + \Delta t\, h_{\mathbb{D}^{-1}}^{-2} \sum_K |K| \left(\frac{h_{\mathbb{D}^{-1}}}{a_{K,\mathbb{D}^{-1}}}\right)^2, & d = 1, \\ (1 + |\ln \psi_{\mathbb{D}}|) \left(1 + \Delta t\, h_{\mathbb{D}^{-1}}^{-2} \sum_K |K| \left(\frac{h_{\mathbb{D}^{-1}}}{a_{K,\mathbb{D}^{-1}}}\right)^2\right), & d = 2, \\ 1 + \Delta t\, h_{\mathbb{D}^{-1}}^{-2} \left(\sum_K |K| \left(\frac{h_{\mathbb{D}^{-1}}}{a_{K,\mathbb{D}^{-1}}}\right)^d\right)^{\frac{2}{d}}, & d \geq 3, \end{cases} \tag{26}$$

$$\psi_{\mathbb{D}} = \frac{1 + \Delta t\, h_{\mathbb{D}^{-1}}^{-2} \max_K \left(\frac{h_{\mathbb{D}^{-1}}}{a_{K,\mathbb{D}^{-1}}}\right)^2}{1 + \Delta t\, h_{\mathbb{D}^{-1}}^{-2} \sum_K \frac{|K|}{|\Omega|} \left(\frac{h_{\mathbb{D}^{-1}}}{a_{K,\mathbb{D}^{-1}}}\right)^2}. \tag{27}$$

Proof. For $P = I_{N_{vi}}$, (24) follows from Lemmas 3.3 and 3.6. For $P = M$, (11) and (13) give

$$\lambda_{\max}(P^{-\frac{1}{2}}AP^{-\frac{1}{2}}) = \max_{\boldsymbol{v}\neq 0} \frac{\boldsymbol{v}^T A \boldsymbol{v}}{\boldsymbol{v}^T P \boldsymbol{v}} \leq 2(d+1) \max_{\boldsymbol{v}\neq 0} \frac{\boldsymbol{v}^T A_D \boldsymbol{v}}{\boldsymbol{v}^T M_D \boldsymbol{v}} = 2(d+1) \max_j \frac{A_{jj}}{M_{jj}},$$

which, together with (11) and (18), implies (24). For $P = M_D$ and $P = M_{lump}$, (24) follows from Lemmas 3.1 and 3.2. For $P = M_D + \Delta t\, A_D$, (24) follows from (13).

The inequality (25) for $P = I_{N_{vi}}$ follows from Lemma 3.7. For $P = M$ and, similarly for $P = M_D$ and $P = M_{lump}$, (25) follows from the basic property of the conformal FE approximation of elliptic eigenvalue problems (e.g.,

Please cite this article as: W. Huang, L. Kamenski and J. Lang, Conditioning of implicit Runge–Kutta integration for finite element approximation of linear diffusion equations on anisotropic meshes, Journal of Computational and Applied Mathematics (2019) 112497, https://doi.org/10.1016/j.cam.2019.112497.

[19, Theorem 1]), since the eigenvalue problem for $P^{-\frac{1}{2}}AP^{-\frac{1}{2}}$ is a FE approximation to the eigenvalue problem for the operator $-\nabla\cdot(\mathbb{D}\nabla)$.

For $P = M_D + \Delta t\, A_D$, using the strategy in the proof of [10, Lemma 5.1], we have

$$\boldsymbol{v}^T P^{-\frac{1}{2}}AP^{-\frac{1}{2}}\boldsymbol{v} \gtrsim d_{\min}\times\begin{cases}\left(\sum_j v_j^2\right)\cdot\left(\sum_j P_{jj}\right)^{-1}, & d=1,\\ \frac{1}{q}\left(\sum_K s_K^{\frac{q}{q-2}}\right)^{-\frac{q-2}{q}}\sum_j v_j^2 P_{jj}^{-1}\sum_{K\in\omega_j} s_K|K|^{\frac{2}{q}}, & d=2,\\ \left(\sum_K s_K^{\frac{d}{2}}\right)^{-\frac{2}{d}}\sum_j v_j^2 P_{jj}^{-1}\sum_{K\in\omega_j} s_K|K|^{\frac{d-2}{d}}, & d\geq 3,\end{cases} \tag{28}$$

with a parameter $q > 2$ and some not-all-zero nonnegative numbers s_K, $K \in \mathcal{T}_h$ (to be determined later). Here, $\sum_K$ denotes the summation over all elements in $\mathcal{T}_h$.

For notational simplicity, we denote $r_K = a^{-2}_{K,\mathbb{D}^{-1}}$ and rewrite (23) as

$$P_{jj} \lesssim \sum_{K\in\omega_j} |K|(1+\Delta t\, r_K). \tag{29}$$

For $d = 1$, (29) implies

$$\boldsymbol{v}^T P^{-\frac{1}{2}}AP^{-\frac{1}{2}}\boldsymbol{v} \gtrsim \frac{d_{\min}\sum_j v_j^2}{\sum_K |K|\,(1+\Delta t\, r_K)}$$

and therefore

$$\lambda_{\min}(P^{-\frac{1}{2}}AP^{-\frac{1}{2}}) \gtrsim \frac{d_{\min}}{\sum_K |K|\,(1+\Delta t\, r_K)}. \tag{30}$$

For $d = 2$, we choose $s_K = |K|^{1-\frac{2}{q}}(1+\Delta t\, r_K)$. Then, (29) implies

$$P_{jj}^{-1}\sum_{K\in\omega_j} s_K|K|^{\frac{2}{q}} \geq C$$

and using (28) we obtain

$$\boldsymbol{v}^T P^{-\frac{1}{2}}AP^{-\frac{1}{2}}\boldsymbol{v} \gtrsim \frac{d_{\min}\sum_k v_k^2}{q\left(\sum_K |K|(1+\Delta t\, r_K)^{\frac{q}{q-2}}\right)^{\frac{q-2}{q}}}.$$

The denominator tends to infinity as $q\to\infty$ and to $2\max_K(1+\Delta t\, r_K)$ as $q\to 2^+$. To find an optimal choice for q, we first estimate the denominator as

$$\begin{aligned} q\left(\sum_K |K|(1+\Delta t\, r_K)^{\frac{q}{q-2}}\right)^{\frac{q-2}{q}} &= q\left(\sum_K |K|(1+\Delta t\, r_K)\cdot(1+\Delta t\, r_K)^{\frac{2}{q-2}}\right)^{\frac{q-2}{q}}\\ &\leq q\left(\left(\sum_K |K|(1+\Delta t\, r_K)\right)\cdot\left(1+\Delta t\,\max_K r_K\right)^{\frac{2}{q-2}}\right)^{\frac{q-2}{q}}\\ &= q\psi_{\mathbb{D}}^{\frac{2}{q}}\sum_K |K|(1+\Delta t\, r_K), \end{aligned}$$

where

$$\psi_{\mathbb{D}} = \frac{1+\Delta t\,\max_K r_K}{\sum_K |K|(1+\Delta t\, r_K)}. \tag{31}$$

If $\psi_{\mathbb{D}} > e$, we choose $q = 2\ln\psi_{\mathbb{D}}$ and obtain $q\psi_{\mathbb{D}}^{\frac{2}{q}} = 2e\ln\psi_{\mathbb{D}}$. If $\psi_{\mathbb{D}} \leq e$, we use $q\to 2^+$ and obtain $q\psi_{\mathbb{D}}^{\frac{2}{q}} \leq 2e$. Combining these two cases yields

$$q\left(\sum_K |K|(1+\Delta t\, r_K)^{\frac{q}{q-2}}\right)^{\frac{q-2}{q}} \leq 2e\,(1+|\ln\psi_{\mathbb{D}}|)\left(\sum_K |K|(1+\Delta t\, r_K)\right)$$

Please cite this article as: W. Huang, L. Kamenski and J. Lang, Conditioning of implicit Runge–Kutta integration for finite element approximation of linear diffusion equations on anisotropic meshes, Journal of Computational and Applied Mathematics (2019) 112497, https://doi.org/10.1016/j.cam.2019.112497.

and therefore

$$\lambda_{\min}(P^{-\frac{1}{2}}AP^{-\frac{1}{2}}) \gtrsim \frac{d_{\min}}{(1+|\ln \psi_{\mathbb{D}}|)\left(\sum_K |K|(1+\Delta t\, r_K)\right)}. \tag{32}$$

For $d \geq 3$, we choose $s_K = |K|^{\frac{2}{d}}(1+\Delta t\, r_K)$ such that $P_{jj}^{-1}\sum_{K\in\omega_j} r_K|K|^{\frac{2}{q}} \geq C$. From (28), we have

$$\boldsymbol{v}^T P^{-\frac{1}{2}}AP^{-\frac{1}{2}}\boldsymbol{u} \gtrsim \frac{d_{\min}\sum_j v_j^2}{\left(\sum_K |K|(1+\Delta t\, r_K)^{\frac{d}{2}}\right)^{\frac{2}{d}}},$$

which gives

$$\lambda_{\min}(P^{-\frac{1}{2}}AP^{-\frac{1}{2}}) \gtrsim \frac{d_{\min}}{\left(\sum_K |K|(1+\Delta t\, r_K)^{\frac{d}{2}}\right)^{\frac{2}{d}}}. \tag{33}$$

Further, (31) can be rewritten into (27) and

$$\sum_K |K|(1+\Delta t\, r_K) = |\Omega| + \Delta t\, h_{\mathbb{D}^{-1}}^{-2}\sum_K |K|\left(\frac{h_{\mathbb{D}^{-1}}}{a_{K,\mathbb{D}^{-1}}}\right)^2,$$

$$\left(\sum_K |K|(1+\Delta t\, r_K)^{\frac{d}{2}}\right)^{\frac{2}{d}} \lesssim |\Omega|^{\frac{2}{d}} + \Delta t\, h_{\mathbb{D}^{-1}}^{-2}\left(\sum_K |K|\left(\frac{h_{\mathbb{D}^{-1}}}{a_{K,\mathbb{D}^{-1}}}\right)^d\right)^{\frac{2}{d}}.$$

Combining these with (30) to (33) gives (25) with $P = M_D + \Delta t\, A_D$. □

4. Conditioning of the implicit Euler integration

In the following we estimate the conditioning of $M + \Delta t\, A$ (in the l_2-norm), which is related to the efficient iterative solution of (9).

4.1. Condition number of $M + \Delta t\, A$

Theorem 4.1. *The condition number of $M + \Delta t\, A$ is bounded by*

$$\kappa(M+\Delta t\, A) \lesssim \frac{\max_j \frac{|\omega_j|}{|\omega_{\min}|}\left(1+\Delta t\, h_{\mathbb{D}^{-1}}^{-2}\sum_{K\in\omega_j}\frac{|K|}{|\omega_j|}\left(\frac{h_{\mathbb{D}^{-1}}}{a_{K,\mathbb{D}^{-1}}}\right)^2\right)}{1+\Delta t\, d_{\min}N^{-1}|\omega_{\min}|^{-1}\Psi_E^{-1}}. \tag{34}$$

Proof. (11) and (13) yield

$$M + \Delta t\, A \leq \frac{d+2}{2}M_D + \Delta t\,(d+1)A_D \leq (d+1)(M_D + \Delta t\, A_D).$$

From this and (11) and (18) we obtain

$$\lambda_{\max}(M+\Delta t\, A) \lesssim \max_j\left(|\omega_j| + \Delta t\, h_{\mathbb{D}^{-1}}^{-2}\sum_{K\in\omega_j}|K|\left(\frac{h_{\mathbb{D}^{-1}}}{a_{K,\mathbb{D}^{-1}}}\right)^2\right).$$

On the other hand, using $\lambda_{\min}(M+\Delta t\, A) \geq \lambda_{\min}(M) + \Delta t\, \lambda_{\min}(A)$ and Lemmas 3.8 and 3.9 we have

$$\lambda_{\min}(M+\Delta t\, A) \gtrsim |\omega_{\min}| + \Delta t\, d_{\min}N^{-1}\Psi_E^{-1}.$$

Combining the above results yields (34). □

There are three factors influencing bound (34). The first factor is the number of the mesh elements N. The second factor is

$$\sum_{K\in\omega_j}\frac{|K|}{|\omega_j|}\left(\frac{h_{\mathbb{D}^{-1}}}{a_{K,\mathbb{D}^{-1}}}\right)^2, \tag{35}$$

Please cite this article as: W. Huang, L. Kamenski and J. Lang, Conditioning of implicit Runge–Kutta integration for finite element approximation of linear diffusion equations on anisotropic meshes, Journal of Computational and Applied Mathematics (2019) 112497, https://doi.org/10.1016/j.cam.2019.112497.

which reflects the mesh nonuniformity in the metric $\mathbb{D}^{-1}$ and is a constant for a $\mathbb{D}^{-1}$-uniform mesh, for which $a_{K,\mathbb{D}^{-1}} \sim h_{\mathbb{D}^{-1}}$ for all K. The third factor is the effect of the mesh nonuniformity (in the Euclidean metric) reflected by Ψ_E, $|\omega_j|/|\omega_{\min}|$, and $N|\omega_{\min}|$, which all become constants if the mesh is uniform.

The time step size Δt plays the role of a homotopy parameter between the mass and the stiffness matrices:

$$\kappa(M + \Delta t\, A) \xrightarrow{\Delta t \to 0} \kappa(M) \lesssim \frac{|\omega_{\max}|}{|\omega_{\min}|}. \tag{36}$$

Thus, $\kappa(M)$ depends on the nonuniformity of the element patch volumes. On the other hand,

$$\kappa(M + \Delta t\, A) \xrightarrow{\Delta t \to \infty} \kappa(A) \lesssim d_{\min}^{-1} h_{\mathbb{D}^{-1}}^{-2} \Psi_E \max_j \left(N \sum_{K \in \omega_j} |K| \left(\frac{h_{\mathbb{D}^{-1}}}{a_{K,\mathbb{D}^{-1}}} \right)^2 \right), \tag{37}$$

which has been obtained previously [10, Theorem 5.2]. It depends on three factors as well: N (through $h_{\mathbb{D}^{-1}}$), the mesh nonuniformity in the Euclidean metric (through Ψ_E), and the mesh nonuniformity with respect to $\mathbb{D}^{-1}$ (through the term in the maximum).

4.2. Condition number of $P^{-\frac{1}{2}}(M + \Delta t\, A)P^{-\frac{1}{2}}$

Theorem 4.2. *The condition number of $P^{-\frac{1}{2}}(M + \Delta t\, A)P^{-\frac{1}{2}}$ with $P = M$, M_D, or M_{lump} is bounded by*

$$\kappa(P^{-\frac{1}{2}}(M + \Delta t\, A)P^{-\frac{1}{2}}) \lesssim \frac{1 + \Delta t\, h_{\mathbb{D}^{-1}}^{-2} \max_j \sum_{K \in \omega_j} \frac{|K|}{|\omega_j|} \left(\frac{h_{\mathbb{D}^{-1}}}{a_{K,\mathbb{D}^{-1}}} \right)^2}{1 + \Delta t\, \lambda_{\mathbb{D}}}, \tag{38}$$

where $\lambda_{\mathbb{D}}$ is the minimal eigenvalue of $-\nabla \cdot (\mathbb{D}\nabla)$.

Proof. Bound (38) is obtained by combining Lemmas 3.8 and 3.9 with the estimates

$$\lambda_{\max}\left(P^{-\frac{1}{2}}(M + \Delta t\, A)P^{-\frac{1}{2}}\right) \le \lambda_{\max}\left(P^{-\frac{1}{2}}MP^{-\frac{1}{2}}\right) + \Delta t\, \lambda_{\max}\left(P^{-\frac{1}{2}}AP^{-\frac{1}{2}}\right),$$
$$\lambda_{\min}\left(P^{-\frac{1}{2}}(M + \Delta t\, A)P^{-\frac{1}{2}}\right) \ge \lambda_{\min}\left(P^{-\frac{1}{2}}MP^{-\frac{1}{2}}\right) + \Delta t\, \lambda_{\min}\left(P^{-\frac{1}{2}}AP^{-\frac{1}{2}}\right). \quad \square$$

The bounds on $\kappa(P^{-\frac{1}{2}}(M + \Delta t\, A)P^{-\frac{1}{2}})$ for $P = M$, M_D, and M_{lump} are similar, while the last two choices lead to a simple diagonal scaling, which is easier to implement.

All three choices reduce the effects of the mesh nonuniformity: in comparison to (34), bound (38) does not depend on $|\omega_{\min}|$ or Ψ_E directly and contains only the $\mathbb{D}^{-1}$-nonuniformity factor (35). This is intuitive, since the eigenvalues of $M^{-1}A$ approximate those of the underlying continuous operator, which are mesh-independent. However, $\kappa(M^{-1}A)$ is not necessarily smaller then $\kappa(A)$ and the overall effect depends on the magnitude of Δt: if Δt is large, $\kappa(M^{-1}(M+\Delta t A))$ might not be better than $\kappa(M + \Delta t A)$. On the other hand, if Δt is small, we can expect that $\kappa(M^{-1}(M + \Delta t A)) < \kappa(M + \Delta t A)$. The numerical experiments in Section 6 support this argument (see Example 6.2 and Fig. 3b).

Theorem 4.3. *The condition number of $P^{-\frac{1}{2}}(M + \Delta t\, A)P^{-\frac{1}{2}}$ with the Jacobi preconditioner $P = M_D + \Delta t\, A_D$ is bounded by*

$$\kappa(P^{-\frac{1}{2}}(M + \Delta t\, A)P^{-\frac{1}{2}}) \lesssim \left(\frac{1}{1 + \Delta t\, h_{\mathbb{D}^{-1}}^{-2} \max_j \sum_{K \in \omega_j} \frac{|K|}{|\omega_j|} \left(\frac{h_{\mathbb{D}^{-1}}}{a_{K,\mathbb{D}^{-1}}} \right)^2} + \frac{d_{\min}}{\Psi_{\mathbb{D}}} \right)^{-1} \tag{39}$$

where $\Psi_{\mathbb{D}}$ is given in (26).

Proof. The proof is similar to that of Theorem 4.2. $\square$

Bound (39) is comparable to (38) although the former is smaller than the latter in general, especially for large Δt, since the factor $\Psi_{\mathbb{D}}$ in (39) involves averaging over all elements, whereas (38) involves the maximum over patch averages.

In (39), Δt is a homotopy parameter between the mass and stiffness matrices:

$$\kappa(P^{-\frac{1}{2}}(M + \Delta t\, A)P^{-\frac{1}{2}}) \xrightarrow{\Delta t \to 0} \kappa(M_D^{-\frac{1}{2}} M M_D^{-\frac{1}{2}}) \le C.$$

Please cite this article as: W. Huang, L. Kamenski and J. Lang, Conditioning of implicit Runge–Kutta integration for finite element approximation of linear diffusion equations on anisotropic meshes, Journal of Computational and Applied Mathematics (2019) 112497, https://doi.org/10.1016/j.cam.2019.112497.

On the other hand,

$$\kappa(P^{-\frac{1}{2}}(M+\Delta t\,A)P^{-\frac{1}{2}}) \xrightarrow{\Delta t\to\infty} \kappa(A_D^{-\frac{1}{2}}AA_D^{-\frac{1}{2}})$$
$$\lesssim \frac{N^{\frac{2}{d}}}{d_{\min}|\Omega|_{\mathbb{D}^{-1}}^{\frac{2}{d}}} \times \begin{cases} \sum_K |K|\left(\frac{h_{\mathbb{D}^{-1}}}{a_{K,\mathbb{D}^{-1}}}\right)^2, & d=1,\\ (1+|\ln\psi_{\mathbb{D}}|)\sum_K |K|\left(\frac{h_{\mathbb{D}^{-1}}}{a_{K,\mathbb{D}^{-1}}}\right)^2, & d=2,\\ \left(\sum_K |K|\left(\frac{h_{\mathbb{D}^{-1}}}{a_{K,\mathbb{D}^{-1}}}\right)^d\right)^{\frac{2}{d}}, & d\ge 3, \end{cases}$$

where $\psi_{\mathbb{D}}$ becomes

$$\psi_{\mathbb{D}} = \frac{\max_K \left(\frac{h_{\mathbb{D}^{-1}}}{a_{K,\mathbb{D}^{-1}}}\right)^2}{\sum_K \frac{|K|}{|\Omega|}\left(\frac{h_{\mathbb{D}^{-1}}}{a_{K,\mathbb{D}^{-1}}}\right)^2}.$$

This bound is equivalent to the bound obtained in [10, Theorem 5.2].

Remark 4.1. In comparison to the bound (34) for $M+\Delta t\,A$, bounds (38) and (39) contain only two mesh-dependent factors: N and the mesh nonuniformity in the metric $\mathbb{D}^{-1}$ (through the terms involving the ratio $h_{\mathbb{D}^{-1}}/a_{K,\mathbb{D}^{-1}}$). This shows that the effects of the mesh nonuniformity (in the Euclidean metric) on the condition number is effectively eliminated by the preconditioning.

Remark 4.2. For a $\mathbb{D}^{-1}$-uniform mesh, $a_{K,\mathbb{D}^{-1}} \sim h_{\mathbb{D}^{-1}}$ and, hence, all terms involving the ratio $h_{\mathbb{D}^{-1}}/a_{K,\mathbb{D}^{-1}}$ will become a constant. Thus,

$$\kappa(P^{-\frac{1}{2}}(M+\Delta t\,A)P^{-\frac{1}{2}}) = \mathcal{O}(1+\Delta t\,N^{\frac{2}{d}}),$$

which shows more clearly the role of Δt:

$$\kappa(P^{-\frac{1}{2}}(M+\Delta t\,A)P^{-\frac{1}{2}}) = \begin{cases} \mathcal{O}(N^{\frac{2}{d}}), & \Delta t = \mathcal{O}(1),\\ \mathcal{O}(N^{\frac{1}{d}}), & \Delta t = \mathcal{O}(h_{\mathbb{D}^{-1}}) = \mathcal{O}(N^{-\frac{1}{d}}),\\ \mathcal{O}(1), & \Delta t = \mathcal{O}(N^{-\frac{2}{d}}) \text{ or } \Delta t \to 0. \end{cases}$$

5. Conditioning of general implicit RK integration

For a general implicit RK method, the matrix $\mathbb{A}$ in (8) is not necessarily normal and in this case a condition number in the standard definition does not provide much information for the convergence of iterative methods. When its symmetric part, $(\mathbb{A}+\mathbb{A}^T)/2$, is positive definite, it is known [14] that the convergence of the generalized minimal residual method (GMRES) is

$$\|r_n\| \le \left(1 - \frac{\lambda_{\min}^2((\mathbb{A}+\mathbb{A}^T)/2)}{\sigma_{\max}^2(\mathbb{A})}\right)^{n/2} \|r_0\|, \tag{40}$$

where $\sigma_{\max}(\mathbb{A})$ is the largest singular value of $\mathbb{A}$, $\lambda_{\min}((\mathbb{A}+\mathbb{A}^T)/2)$ is the minimal eigenvalue of the symmetric part, r_n is the residual of the corresponding linear system at the nth iterate, and $\|\cdot\|$ stands for the matrix or vector 2-norm. Thus, we can consider the "condition number"

$$\tilde{\kappa}(\mathbb{A}) = \frac{\sigma_{\max}(\mathbb{A})}{\lambda_{\min}((\mathbb{A}+\mathbb{A}^T)/2)}. \tag{41}$$

This definition reduces to the standard definition of the condition number (in 2-norm) for symmetric matrices.

We first recall that for an arbitrary square matrix U, its maximal singular value is given by

$$\sigma_{\max}(U) = \max_{\boldsymbol{v}\neq 0} \frac{\|U\boldsymbol{v}\|}{\|\boldsymbol{v}\|}.$$

From this, for any square matrices U and V we have

$$\sigma_{\max}(U+V) \le \sigma_{\max}(U) + \sigma_{\max}(V). \tag{42}$$

The next lemma provides the eigenvalues and singular values of the Kronecker product of any square matrices U and V.

Please cite this article as: W. Huang, L. Kamenski and J. Lang, Conditioning of implicit Runge–Kutta integration for finite element approximation of linear diffusion equations on anisotropic meshes, Journal of Computational and Applied Mathematics (2019) 112497, https://doi.org/10.1016/j.cam.2019.112497.

Lemma 5.1 ([17, Theorems 13.10 and 13.12]). *For any square matrices U and V, the eigenvalues and singular values of $U \otimes V$ are $\lambda_j(U)\lambda_k(V)$ and $\sigma_j(U)\sigma_k(V)$, $j, k = 1, 2, \ldots$*

The above lemma in particular implies

$$\sigma_{\max}(U \otimes V) = \sigma_{\max}(U)\sigma_{\max}(V)$$

and, for a positive semi-definite matrix V,

$$\lambda_{\min}(U \otimes V) = \lambda_{\min}(U) \times \begin{cases} \lambda_{\min}(V), & \text{if } \lambda_{\min}(U) \geq 0, \\ \lambda_{\max}(V), & \text{if } \lambda_{\min}(U) < 0. \end{cases} \tag{43}$$

5.1. General implicit RK methods: simultaneous solution

Although a successive strategy for implicit RK methods is more common (e.g., [20, p. 131]), the simultaneous solution can be appealing if an iterative method is used. It is also important to have a theoretical understanding of the overall system conditioning.

Theorem 5.1. *Let P be a symmetric and positive definite preconditioner, Γ an $s \times s$ implicit RK coefficient matrix, and $\mathbb{A} = I_s \otimes M + \Delta t\, \Gamma A$.*

If $\lambda_{\min}((\Gamma + \Gamma^T)/2) < 0$ and Δt is chosen sufficiently small such that

$$\lambda_{\min}(P^{-\frac{1}{2}} M P^{-\frac{1}{2}}) + \Delta t \lambda_{\min}((\Gamma + \Gamma^T)/2)\lambda_{\max}(P^{-\frac{1}{2}} A P^{-\frac{1}{2}}) > 0, \tag{44}$$

then

$$\tilde{\kappa}\left((I_s \otimes P^{-\frac{1}{2}})\mathbb{A}(I_s \otimes P^{-\frac{1}{2}})\right) \leq \frac{\lambda_{\max}(P^{-\frac{1}{2}} M P^{-\frac{1}{2}}) + \Delta t\, \sigma_{\max}(\Gamma)\lambda_{\max}(P^{-\frac{1}{2}} A P^{-\frac{1}{2}})}{\lambda_{\min}(P^{-\frac{1}{2}} M P^{-\frac{1}{2}}) + \Delta t \lambda_{\min}((\Gamma + \Gamma^T)/2)\lambda_{\max}(P^{-\frac{1}{2}} A P^{-\frac{1}{2}})}. \tag{45}$$

If $\lambda_{\min}((\Gamma + \Gamma^T)/2) \geq 0$, then

$$\tilde{\kappa}\left((I_s \otimes P^{-\frac{1}{2}})\mathbb{A}(I_s \otimes P^{-\frac{1}{2}})\right) \leq \frac{\lambda_{\max}(P^{-\frac{1}{2}} M P^{-\frac{1}{2}}) + \Delta t\, \sigma_{\max}(\Gamma)\lambda_{\max}(P^{-\frac{1}{2}} A P^{-\frac{1}{2}})}{\lambda_{\min}(P^{-\frac{1}{2}} M P^{-\frac{1}{2}}) + \Delta t \lambda_{\min}((\Gamma + \Gamma^T)/2)\lambda_{\min}(P^{-\frac{1}{2}} A P^{-\frac{1}{2}})}. \tag{46}$$

Proof. First, notice that

$$(I_s \otimes P^{-\frac{1}{2}})(I_s \otimes M + \Delta t\, \Gamma \otimes A)(I_s \otimes P^{-\frac{1}{2}}) = I_s \otimes (P^{-\frac{1}{2}} M P^{-\frac{1}{2}}) + \Delta t\, \Gamma \otimes (P^{-\frac{1}{2}} A P^{-\frac{1}{2}}).$$

Then, from (42) and Lemma 5.1 we have

$$\begin{aligned} &\sigma_{\max}\left((I_s \otimes P^{-\frac{1}{2}})(I_s \otimes M + \Delta t\, \Gamma \otimes A)(I_s \otimes P^{-\frac{1}{2}})\right) \\ &\quad \leq \sigma_{\max}\left(I_s \otimes (P^{-\frac{1}{2}} M P^{-\frac{1}{2}})\right) + \Delta t\, \sigma_{\max}\left(\Gamma \otimes (P^{-\frac{1}{2}} A P^{-\frac{1}{2}})\right) \\ &\quad = \sigma_{\max}\left(P^{-\frac{1}{2}} M P^{-\frac{1}{2}}\right) + \Delta t\, \sigma_{\max}(\Gamma)\, \sigma_{\max}\left(P^{-\frac{1}{2}} A P^{-\frac{1}{2}}\right) \\ &\quad = \lambda_{\max}\left(P^{-\frac{1}{2}} M P^{-\frac{1}{2}}\right) + \Delta t\, \sigma_{\max}(\Gamma)\, \lambda_{\max}\left(P^{-\frac{1}{2}} A P^{-\frac{1}{2}}\right). \end{aligned} \tag{47}$$

Moreover, the symmetric part of $(I_s \otimes P^{-\frac{1}{2}})(I_s \otimes M + \Delta t\, \Gamma \otimes A)(I_s \otimes P^{-\frac{1}{2}})$ is

$$I_s \otimes (P^{-\frac{1}{2}} M P^{-\frac{1}{2}}) + \Delta t\, (\Gamma + \Gamma^T)/2 \otimes (P^{-\frac{1}{2}} A P^{-\frac{1}{2}}).$$

Then,

$$\begin{aligned} &\lambda_{\min}\left(I_s \otimes (P^{-\frac{1}{2}} M P^{-\frac{1}{2}}) + \Delta t\, (\Gamma + \Gamma^T)/2 \otimes (P^{-\frac{1}{2}} A P^{-\frac{1}{2}})\right) \\ &\geq \lambda_{\min}\left(I_s \otimes (P^{-\frac{1}{2}} M P^{-\frac{1}{2}})\right) + \Delta t\, \lambda_{\min}\left((\Gamma + \Gamma^T)/2 \otimes (P^{-\frac{1}{2}} A P^{-\frac{1}{2}})\right). \end{aligned}$$

In case of $\lambda_{\min}((\Gamma + \Gamma^T)/2) \geq 0$ or if $\lambda_{\min}((\Gamma + \Gamma^T)/2) < 0$ and Δt is chosen such that (44) holds, then the symmetric part of $(I_s \otimes P^{-\frac{1}{2}})(I_s \otimes M + \Delta t\, \Gamma \otimes A)(I_s \otimes P^{-\frac{1}{2}})$ is positive definite and (45) and (46) follow from the above results and (41) and (43). □

Please cite this article as: W. Huang, L. Kamenski and J. Lang, Conditioning of implicit Runge–Kutta integration for finite element approximation of linear diffusion equations on anisotropic meshes, Journal of Computational and Applied Mathematics (2019) 112497, https://doi.org/10.1016/j.cam.2019.112497.

Remark 5.1. In case of $P = I_{N_{vi}}$, M, M_D, M_{lump}, and $M_D + \Delta t\, A_D$, the estimates for the eigenvalues $\lambda_{\min}(P^{-\frac{1}{2}} M P^{-\frac{1}{2}})$, $\lambda_{\max}(P^{-\frac{1}{2}} M P^{-\frac{1}{2}})$, $\lambda_{\min}(P^{-\frac{1}{2}} A P^{-\frac{1}{2}})$, and $\lambda_{\max}(P^{-\frac{1}{2}} A P^{-\frac{1}{2}})$ are given by Lemmas 3.8 and 3.9. The quantities $\sigma_{\max}(\Gamma)$ and $\lambda_{\min}((\Gamma + \Gamma^T)/2)$ are given in Table 1 for five implicit Runge–Kutta methods. One can see that $\sigma_{\max}(\Gamma) = \mathcal{O}(1)$ but $\lambda_{\min}((\Gamma + \Gamma^T)/2)$ can be zero, positive, and negative.

For the 4th-order Gauss and 3rd-order Radau IA methods, $\lambda_{\min}((\Gamma + \Gamma^T)/2) \geq 0$ and (44) poses no constraint on the choice of Δt. (46) indicates that the conditioning of the system (8) resulting from implicit RK integration behaves more or less like that of $M + \Delta t\, A$ for the implicit Euler method, and the bound also involves three mesh-dependent factors and one of them can be eliminated effectively by diagonal preconditioning (see Remark 4.1).

For the other three methods in Table 1 with $\lambda_{\min}((\Gamma + \Gamma^T)/2) < 0$, (44) gives an upper bound on possible Δt. Note that this condition can be serious and lead to a condition like $\Delta t = \mathcal{O}(h^2)$.

5.2. Diagonally implicit RK (DIRK) methods

The coefficient matrix $\Gamma = (\gamma_{kj})_{k,j=1}^{s}$ of a DIRK method is a lower triangular matrix and the system (6) is solved by successively solving s linear systems with $M + \Delta t\, \gamma_{jj} A$, $j = 1, \ldots, s$. The conditioning of $M + \Delta t\, \gamma_{jj} A$ is therefore similar to that of $M + \Delta t\, A$ and, hence, the analysis in Section 4 also applies to DIRK methods.

5.3. General implicit RK methods: successive solution

A successive solution procedure transforms the large system (6) into a number of smaller systems, which are then solved successively. We adopt the approach of Butcher [16] and Bickart [15] and carry out the transformation using the Jordan normal form of the RK matrix. To keep our analysis applicable to methods with a singular Γ, we use the Jordan normal form of Γ instead of Γ^{-1}, which is the conventional choice (e.g., see [20, p. 131]).

Let the Jordan normal form of Γ be

$$\Gamma = T \begin{bmatrix} J_1 & & \\ & \ddots & \\ & & J_p \end{bmatrix} T^{-1}, \tag{48}$$

where T is a real invertible matrix and J_j, $j = 1, \ldots, p$, are the Jordan blocks, which either have the form

$$J_j = \begin{bmatrix} \mu_j & 1 & & \\ & \mu_j & \ddots & \\ & & \ddots & 1 \\ & & & \mu_j \end{bmatrix}, \quad \mu_j \in \mathbb{R}, \tag{49}$$

or

$$J_j = \begin{bmatrix} C_j & I_2 & & \\ & C_j & \ddots & \\ & & \ddots & I_2 \\ & & & C_j \end{bmatrix} \quad \text{with} \quad C_j = \begin{bmatrix} \alpha_j & \beta_j \\ -\beta_j & \alpha_j \end{bmatrix}, \quad \alpha_j, \beta_j \in \mathbb{R}. \tag{50}$$

Recall that the eigenvalues of Γ are assumed to have nonnegative real parts, i.e., $\mu_j, \alpha_j \geq 0$.

Theorem 5.2. *Assume that the eigenvalues of the coefficient matrix Γ for a given implicit RK method have nonnegative real parts and the system (6) is solved by using the Jordan normal form of Γ to transform it into smaller systems. Then, the conditioning of implicit RK integration of (4) is determined by the conditioning of*

$$M + \mu_j \Delta t\, A \quad \text{and} \quad I_2 \otimes M + \Delta t\, C_j \otimes A, \tag{51}$$

where μ_j and $\alpha_j \pm i\beta_j$ are the real and complex eigenvalues of Γ, respectively.

Moreover, for any symmetric and positive definite preconditioner P,

$$\tilde{\kappa}\left((I_2 \otimes P^{-\frac{1}{2}})(I_2 \otimes M + \Delta t\, C_j \otimes A)(I_2 \otimes P^{-\frac{1}{2}})\right) \leq \frac{\lambda_{\max}(P^{-\frac{1}{2}} M P^{-\frac{1}{2}}) + \Delta t \sqrt{\alpha_j^2 + \beta_j^2}\, \lambda_{\max}(P^{-\frac{1}{2}} A P^{-\frac{1}{2}})}{\lambda_{\min}(P^{-\frac{1}{2}} M P^{-\frac{1}{2}}) + \Delta t\, \alpha_j\, \lambda_{\min}(P^{-\frac{1}{2}} A P^{-\frac{1}{2}})}. \tag{52}$$

Please cite this article as: W. Huang, L. Kamenski and J. Lang, Conditioning of implicit Runge–Kutta integration for finite element approximation of linear diffusion equations on anisotropic meshes, Journal of Computational and Applied Mathematics (2019) 112497, https://doi.org/10.1016/j.cam.2019.112497.

Proof. Using (48), we can rewrite (8) as

$$\mathcal{A} = \left(T \otimes I_{N_{vi}}\right)\left(I_s \otimes M + \Delta t \begin{bmatrix} J_1 & & \\ & \ddots & \\ & & J_p \end{bmatrix} \otimes A\right)\left(T^{-1} \otimes I_{N_{vi}}\right)$$
$$= \left(T \otimes I_{N_{vi}}\right)\begin{bmatrix} I_{n_1} \otimes M + \Delta t J_1 \otimes A & & \\ & \ddots & \\ & & I_{n_p} \otimes M + \Delta t J_p \otimes A \end{bmatrix}\left(T^{-1} \otimes I_{N_{vi}}\right),$$

where n_j is the size of J_j (with $\sum_{j=1}^{p} n_j = s$). Since the systems with the coefficient matrix $T \otimes I_{N_{vi}}$ or $T^{-1} \otimes I_{N_{vi}}$ can be solved directly and efficiently, we only need to consider the iterative solution of the systems associated with

$$I_{n_j} \otimes M + \Delta t J_j \otimes A, \quad j = 1, \ldots, p. \tag{53}$$

For a Jordan block in the form (49), the above matrix becomes

$$I_{n_j} \otimes M + \Delta t J_j \otimes A = \begin{bmatrix} M + \mu_j \Delta t A & \Delta t A & & \\ & M + \mu_j \Delta t A & \ddots & \\ & & \ddots & \Delta t A \\ & & & M + \mu_j \Delta t A \end{bmatrix},$$

which can be solved by successively solving n_j systems with $M + \mu_j \Delta t A$ (backward substitution).

On the other hand, for a Jordan block in the form (50), the matrix (53) becomes

$$I_{n_j} \otimes M + \Delta t J_j \otimes A$$
$$= \begin{bmatrix} I_2 \otimes M + \Delta t C_j \otimes A & \Delta t I_2 \otimes A & & \\ & I_2 \otimes M + \Delta t C_j \otimes A & \ddots & \\ & & \ddots & \Delta t I_2 \otimes A \\ & & & I_2 \otimes M + \Delta t C_j \otimes A \end{bmatrix},$$

which, again, can be solved by successively solving $n_j/2$ systems with $I_2 \otimes M + \Delta t C_j \otimes A$. Hence, the conditioning in this case is determined by the matrices in (51).

The analysis of Section 4 can be used to estimate the condition number of $M + \mu_j \Delta t A$. Estimate (52) for the second matrix in (51) is obtained similarly as for Theorem 5.1, except that we now have $\sigma_{\max}(C_j) = \sqrt{\alpha_j^2 + \beta_j^2}$ and $\lambda_{\min}((C_j + C_j^T)/2) = \alpha_j$. □

Note that Remark 5.1 applies to the current situation as well. However, estimate (52) can now be used for all methods in Table 1 *without any restriction on* Δt.

6. Numerical examples

In the following examples we consider the IBVP (1) in 2d ($d = 2$) with $\Omega = (0, 1)^2$ and homogeneous Dirichlet boundary conditions. For the time integration we choose Radau IA method of order 3 with fixed time steps $\Delta t = 10^{-1}$, 10^{-3}, 10^{-5}, and $\Delta t = N^{-\frac{1}{d}} \sim h$, where h is the average mesh size. The Radau IA coefficient matrix is given by

$$\Gamma = \frac{1}{12}\begin{pmatrix} rr3 & -3 \\ 3 & 5 \end{pmatrix} \quad \text{with } \sigma_{\max}(\Gamma) = 0.5 \text{ and } \lambda_{\min}((\Gamma + \Gamma^T)/2) = 0.25.$$

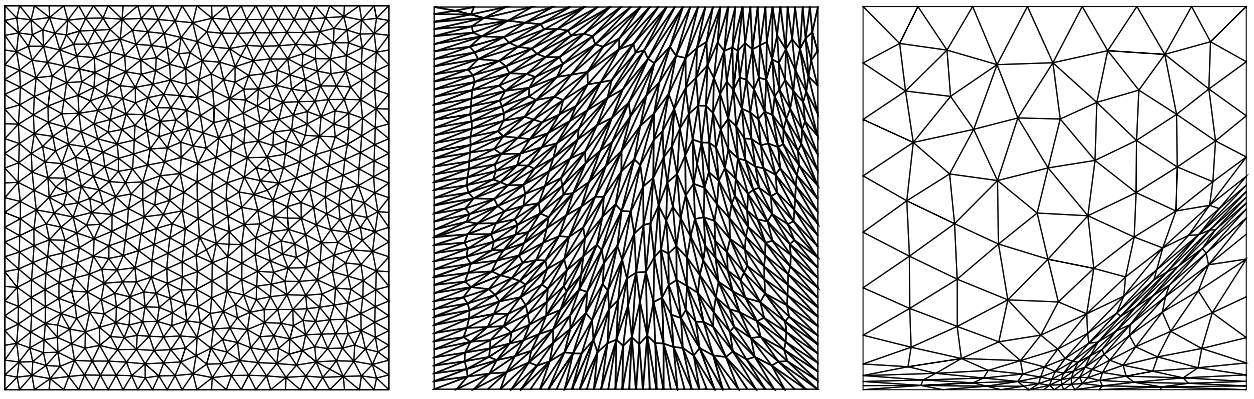
Example 6.1. To compare the condition number $\tilde{\kappa}(I_s \otimes M + \Delta t\, \Gamma \otimes A)$ of Radau IA with the condition number $\kappa(M + \Delta t A)$ of the implicit Euler method we consider two diffusion matrices: isotropic

$$\mathbb{D} = I \quad \text{(Laplace operator)}$$

and anisotropic

$$\mathbb{D}(x, y) = \begin{bmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{bmatrix}\begin{bmatrix} 10 & 0 \\ 0 & 0.1 \end{bmatrix}\begin{bmatrix} \cos\theta & \sin\theta \\ -\sin\theta & \cos\theta \end{bmatrix}, \quad \theta = \pi \sin x \cos y, \tag{54}$$

Please cite this article as: W. Huang, L. Kamenski and J. Lang, Conditioning of implicit Runge–Kutta integration for finite element approximation of linear diffusion equations on anisotropic meshes, Journal of Computational and Applied Mathematics (2019) 112497, https://doi.org/10.1016/j.cam.2019.112497.



(a) quasi-uniform (b) $\mathbb{D}^{-1}$-uniform for (54) (c) adaptive for (55)

Fig. 1. Mesh examples.

with quasi-uniform (Fig. 1a) and $\mathbb{D}^{-1}$-uniform (Fig. 1b) meshes obtained using the mesh generator *bamg* [21]. A quasi-uniform mesh can be also seen as $\mathbb{D}^{-1}$-uniform for $\mathbb{D} = I$.

We expect the conditioning of Radau IA to be similar to that of the implicit Euler (see (45) and Remark 5.1). This is verified in Fig. 2 (left column), which shows that the conditioning for the Radau IA method has the same general behaviour as that for the implicit Euler scheme. With increasing number of mesh elements, the order of conditioning is $\mathcal{O}(N^{\frac{2}{d}})$ for a fixed Δt and becomes $\mathcal{O}(N^{\frac{1}{d}})$ for $\Delta t \sim h \sim N^{-\frac{1}{d}}$ for both methods, which is in perfect agreement with the theoretical prediction in Remark 4.2.

Adapting towards the diffusion matrix improves the conditioning of the stiffness matrix [10]. This effect is observed in our test case as well: for anisotropic diffusion, the conditioning for diffusion-adapted meshes (Fig. 2c) is noticeably smaller than that with quasi-uniform meshes (Fig. 2b), especially as N is getting larger.

To compare the exact values for the conditioning of Radau IA with the bound (45) in Theorem 5.1, we use the exact eigenvalues for the matrices $P^{-1/2}MP^{-1/2}$ and $P^{-1/2}AP^{-1/2}$. Fig. 2 (right column) shows that the bound (46) is greater than the exact condition number but exhibits essentially the same behaviour as N increases.

Example 6.2. To illustrate the effect of the diagonal scaling, we consider $\mathbb{D} = I$ (Laplace operator), $P = M_D + \Delta t\, A_D$, M_D and M_{lump}, and adaptive anisotropic meshes for the interpolation of the function

$$u(x, y) = \tanh(60y) - \tanh(60(x - y - 0.5)), \quad (x, y) \in (0, 1)^2, \tag{55}$$

generated with *bamg* [21] using an anisotropic adaptive metric [22]. This function simulates the interaction between a boundary layer along the x-axis and a shock wave along the line $x = y + 0.5$ (see Fig. 1c for a mesh example).

Consistently with Remark 4.1, Fig. 3a shows that the diagonal scaling with $P = M_D + \Delta t\, A_D$ (Euler) and $I_s \otimes M_D + \Delta t\, \Gamma \otimes A_D$ (Radau IA) reduces the effects caused by the mesh nonuniformity: the conditioning of both Euler and Radau IA methods is reduced by a factor ranging from 3 ($\Delta t = 10^{-1}$) to 8 ($\Delta t = 10^{-3},\ 10^{-5}$).

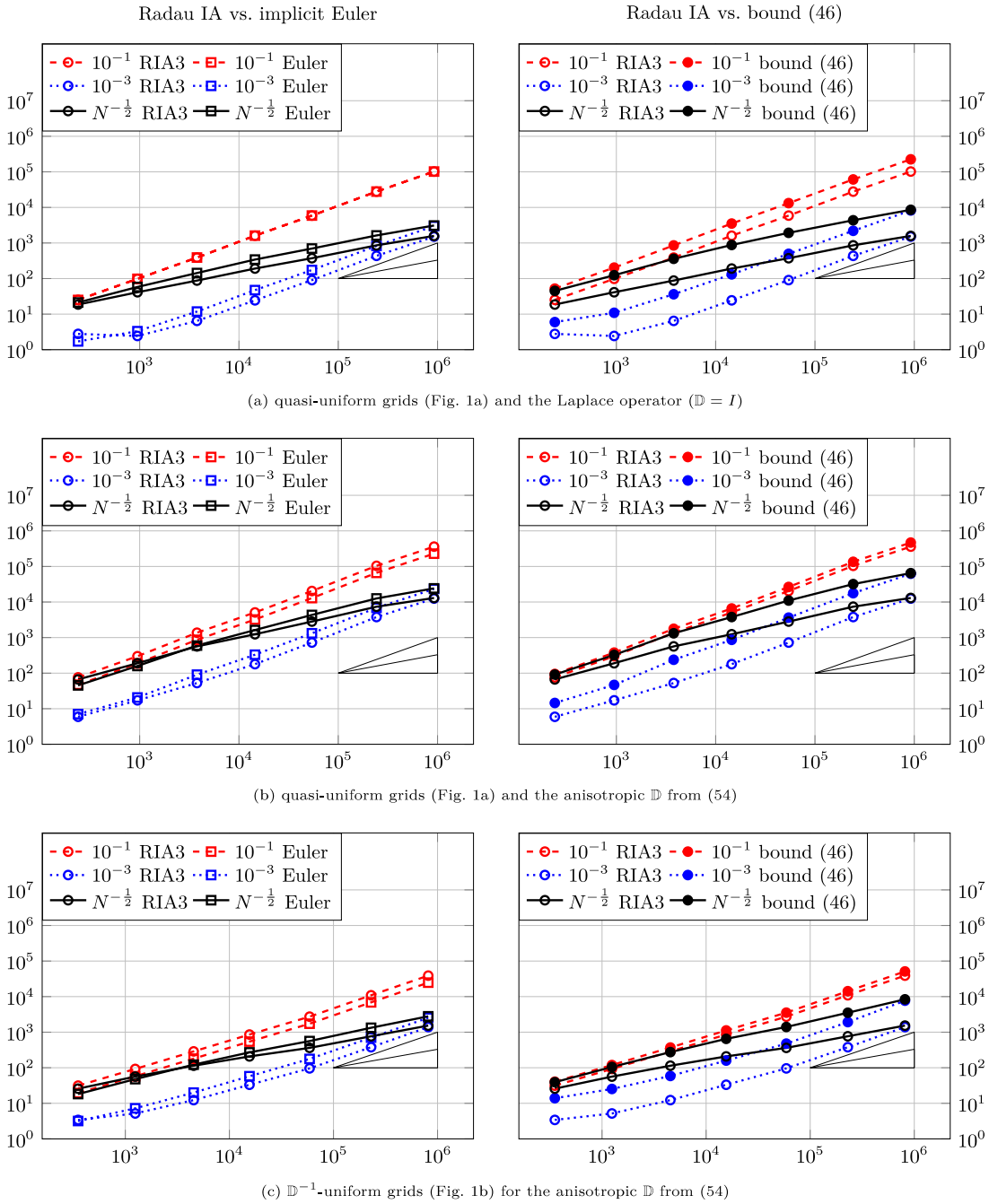
For the scaling with $P = M_D$, Fig. 3b shows that the conditioning is getting worse for large Δt ($\Delta t = 10^{-1}$) but is improving for small Δt ($\Delta t = 10^{-3}$, $\Delta t = 10^{-5}$). To explain this, we recall that the diagonal entries of M are proportional to the corresponding patch volumes (see Lemma 3.1). Hence, scaling with $\mathbb{M}_D$ improves the conditioning issues caused by the mesh *volume-nonuniformity* (the same applies to M and M_{lump}). On the other hand, $\max_j A_{jj} \le \lambda_{\max}(A) \le (d+1)\max_j A_{jj}$ [10, Lemma 4.1] and Lemma 3.6 imply that the largest eigenvalue of A depends on the element shape (see also [10, section 4.1] and [8, section 3]). Thus, $\kappa(M^{-1}A)$ is not necessarily smaller than $\kappa(A)$ since rescaling of A with respect to the patch volumes does not necessarily improve the conditioning issues caused by the shape. Therefore, in general, the overall effect depends on the magnitude of Δt: $\kappa(I + \Delta t M^{-1}A)$ is not necessarily better than $\kappa(M + \Delta t A)$ for large Δt, but we can expect that $\kappa(I + \Delta t M^{-1}A) < \kappa(M + \Delta t A)$ for small Δt.

7. Conclusions

For the FE equations, the conditioning of implicit RK methods with positive semidefinite symmetric part of the coefficient matrix is comparable to the implicit Euler integration (Theorem 5.1).

If the smallest eigenvalue of the symmetric part of the coefficient matrix is negative, (44) gives an upper bound on possible Δt so that the system matrix $I_s \otimes M + \Delta t\, \Gamma \otimes A$ is positive definite. This condition implies $\Delta t \le C \frac{\lambda_{\min}(B)}{\lambda_{\max}(A)}$ and can lead to a condition $\Delta t = \mathcal{O}(h^2)$.

Please cite this article as: W. Huang, L. Kamenski and J. Lang, Conditioning of implicit Runge–Kutta integration for finite element approximation of linear diffusion equations on anisotropic meshes, Journal of Computational and Applied Mathematics (2019) 112497, https://doi.org/10.1016/j.cam.2019.112497.



(a) quasi-uniform grids (Fig. 1a) and the Laplace operator ($\mathbb{D} = I$)

(b) quasi-uniform grids (Fig. 1a) and the anisotropic $\mathbb{D}$ from (54)

(c) $\mathbb{D}^{-1}$-uniform grids (Fig. 1b) for the anisotropic $\mathbb{D}$ from (54)

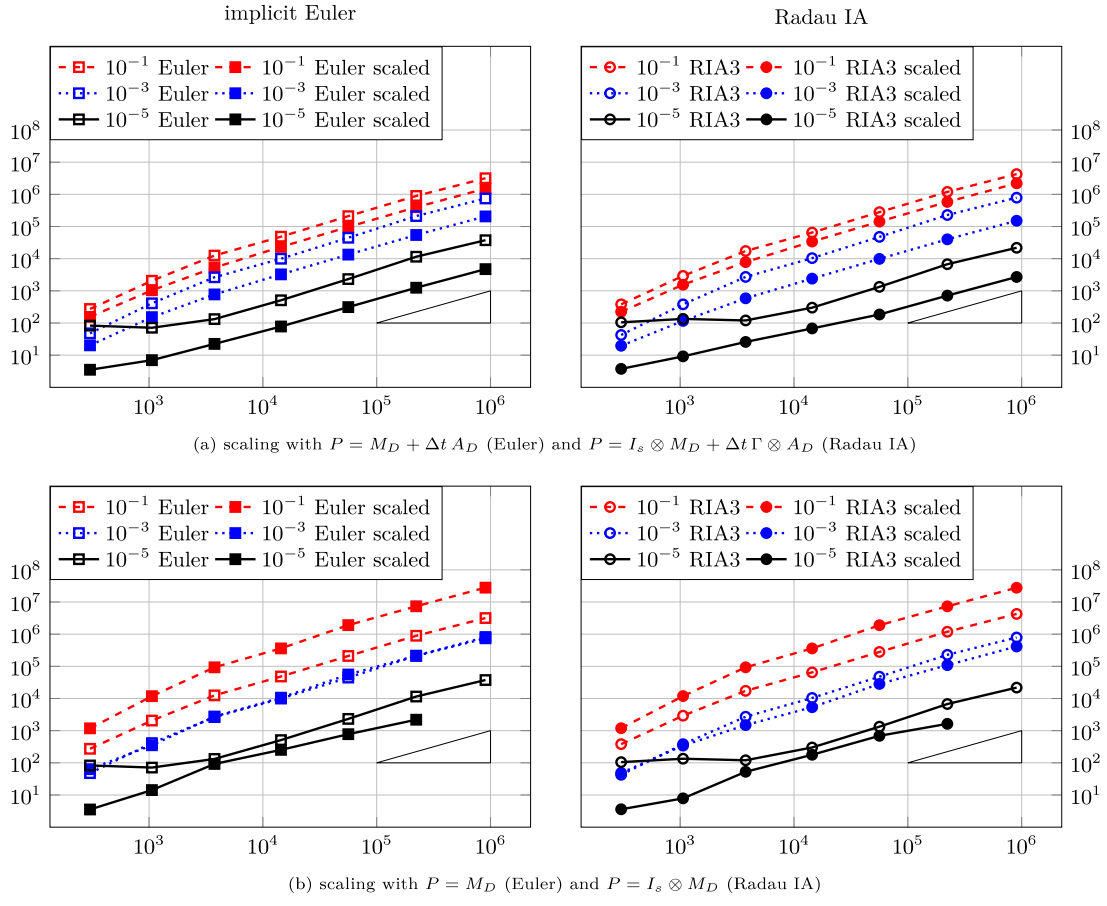
Fig. 2. Conditioning for Radau IA (RIA3) and implicit Euler (left) and a comparison of Radau IA conditioning with the bound (46) (right) as functions of N (Example 6.1).

For the successive solution procedure, the conditioning of the system matrix $I_s \otimes M + \Delta t\, \Gamma \otimes A$ of an implicit RK integration is determined by two types of smaller matrices (Theorem 5.2)

$$M + \mu_j \Delta t\, A \quad \text{and} \quad I_2 \otimes M + \Delta t \begin{bmatrix} \alpha_j & \beta_j \\ -\beta_j & \alpha_j \end{bmatrix} \otimes A.$$

The first matrix is similar to the implicit Euler method and corresponds to the real eigenvalues μ_j of the RK matrix Γ while the second corresponds to the complex eigenvalues $\alpha_j \pm i\beta_j$ of Γ.

Please cite this article as: W. Huang, L. Kamenski and J. Lang, Conditioning of implicit Runge–Kutta integration for finite element approximation of linear diffusion equations on anisotropic meshes, Journal of Computational and Applied Mathematics (2019) 112497, https://doi.org/10.1016/j.cam.2019.112497.



(a) scaling with $P = M_D + \Delta t\, A_D$ (Euler) and $P = I_s \otimes M_D + \Delta t\, \Gamma \otimes A_D$ (Radau IA)

(b) scaling with $P = M_D$ (Euler) and $P = I_s \otimes M_D$ (Radau IA)

Fig. 3. Conditioning of implicit Euler and Radau IA (RIA3) as function of N before and after a diagonal scaling for the Laplace operator and adaptive anisotropic meshes (Example 6.2).

There are three mesh-dependent factors that affect the conditioning of general implicit RK methods: the number of mesh elements (average mesh size), the mesh nonuniformity (in the Euclidean metric), and the mesh nonuniformity with respect to the inverse of the diffusion matrix. Preconditioning by the diagonal part of the system matrix itself or the mass matrix or by the lumped mass matrix reduces the effects of the mesh nonuniformity in the Euclidean metric (Theorems 4.2 and 4.3). These results are consistent with previous studies for the boundary value problems [2–4,10] and explicit integration of linear diffusion problems [11,12].

Similar result can be shown for higher order FE approximations. The major result is based on the preliminary estimates for the mass and stiffness matrices (Lemmas 3.1–3.4 and 3.7) which, in their essential part, hold for higher order as well (with different constants depending of the order). For the FE, a symmetric positive definite (possibly lumped) mass matrix is equivalent to a diagonal matrix of node patch volumes [11, Lemmas 2.4] so that estimates equivalent to Lemmas 3.1 and 3.2 hold for higher order as well (with different constants). For the stiffness matrix, estimates equivalent to Lemmas 3.3 and 3.4 are given in [11, Lemmas 2.2 and 2.3] and the derivation of the bound on $\lambda_{\min}(A)$ in Lemma 3.7 does not depend on the FE order (up to a constant) and therefore holds also for higher order FE.

References

[1] I. Fried, Bounds on the spectral and maximum norms of the finite element stiffness, flexibility and mass matrices, Int. J. Solids Struct. 9 (1973) 1013–1034.
[2] R.E. Bank, L.R. Scott, On the conditioning of finite element equations with highly refined meshes, SIAM J. Numer. Anal. 26 (6) (1989) 1383–1394.
[3] M. Ainsworth, W. McLean, T. Tran, The conditioning of boundary element equations on locally refined meshes and preconditioning by diagonal scaling, SIAM J. Numer. Anal. 36 (6) (1999) 1901–1932.
[4] I.G. Graham, W. McLean, Anisotropic mesh refinement: the conditioning of Galerkin boundary element matrices and simple preconditioners, SIAM J. Numer. Anal. 44 (4) (2006) 1487–1513.
[5] Q. Du, D. Wang, L. Zhu, On mesh geometry and stiffness matrix conditioning for general finite element spaces, SIAM J. Numer. Anal. 47 (2) (2009) 1421–1444.

Please cite this article as: W. Huang, L. Kamenski and J. Lang, Conditioning of implicit Runge–Kutta integration for finite element approximation of linear diffusion equations on anisotropic meshes, Journal of Computational and Applied Mathematics (2019) 112497, https://doi.org/10.1016/j.cam.2019.112497.

[6] L. Zhu, Q. Du, Mesh-dependent stability for finite element approximations of parabolic equations with mass lumping, J. Comput. Appl. Math. 236 (5) (2011) 801–811.
[7] L. Zhu, Q. Du, Mesh dependent stability and condition number estimates for finite element approximations of parabolic problems, Math. Comp. 83 (285) (2014) 37–64.
[8] J.R. Shewchuk, What is a good linear element? Interpolation, conditioning, and quality measures, in: Proceedings of the 11th International Meshing Roundtable, Sandia National Laboratories, 2002, pp. 115–126.
[9] L. Kamenski, W. Huang, A study on the conditioning of finite element equations with arbitrary anisotropic meshes via a density function approach, J. Math. Study 47 (2) (2014) 151–172, arXiv:13026868.
[10] L. Kamenski, W. Huang, H. Xu, Conditioning of finite element equations with arbitrary anisotropic meshes, Math. Comp. 83 (2014) 2187–2211, arXiv:12013651.
[11] W. Huang, L. Kamenski, J. Lang, Stability of explicit runge-kutta methods for high order finite element approximation of linear parabolic equations, in: A. Abdulle, S. Deparis, D. Kressner, F. Nobile, M. Picasso (Eds.), Numerical Mathematics and Advanced Applications — ENUMATH 2013, in: Lect. Notes Comput. Sci. Eng., vol. 103, Springer, Cham, 2015, pp. 165–173, arXiv:190805374.
[12] W. Huang, L. Kamenski, J. Lang, Stability of explicit one-step methods for P1-finite element approximation of linear diffusion equations on anisotropic meshes, SIAM J. Numer. Anal. 54 (3) (2016) 1612–1634, arXiv:160208055.
[13] N.J. Higham, Accuracy and Stability of Numerical Algorithms, Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA, 1996.
[14] S.C. Eisenstat, H.C. Elman, M.H. Schultz, Variational iterative methods for nonsymmetric systems of linear equations, SIAM J. Numer. Anal. 20 (2) (1983) 345–357.
[15] T.A. Bickart, An efficient solution process for implicit Runge-Kutta methods, SIAM J. Numer. Anal. 14 (6) (1977) 1022–1027.
[16] J.C. Butcher, On the implementation of implicit Runge-Kutta methods, Nord. Tidskr. Inf.behandl. (BIT) 16 (3) (1976) 237–240.
[17] A.J. Laub, Matrix Analysis for Scientists & Engineers, Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA, 2005.
[18] W. Huang, L. Kamenski, On the mesh nonsingularity of the moving mesh PDE method, Math. Comp. 87 (312) (2017) 1887–1911, arXiv:151204971.
[19] G.J. Fix, Eigenvalue approximation by the finite element method, Adv. Math. 10 (1973) 300–316.
[20] E. Hairer, G. Wanner, Solving Ordinary Differential Equations. II, second ed., in: Springer Series in Computational Mathematics, vol. 14, Springer-Verlag, Berlin, 1996.
[21] F. Hecht, BAMG: Bidimensional Anisotropic Mesh Generator, URL https://www.ljll.math.upmc.fr/hecht/ftp/bamg, 2006.
[22] W. Huang, Mathematical principles of anisotropic mesh adaptation, Commun. Comput. Phys. 1 (2) (2006) 276–310.

Please cite this article as: W. Huang, L. Kamenski and J. Lang, Conditioning of implicit Runge–Kutta integration for finite element approximation of linear diffusion equations on anisotropic meshes, Journal of Computational and Applied Mathematics (2019) 112497, https://doi.org/10.1016/j.cam.2019.112497.