Aligned numerical methods for anisotropic elliptic problems in bounded domains for plasma edge simulations

Juan Antonio Soler Vasco

Thèse

présentée par

Juan Antonio Soler Vasco

pour l'obtention du

Doctorat de l'Ecole Centrale Marseille

Spécialité: **Fusion magnétique**

Méthodes numériques alignées pour problèmes elliptiques anisotropes en domaines bornés pour simulations du plasma de bord

Thèse soutenue publiquement le 20 Septembre 2019 devant le jury composé de:

Rapporteurs:	Bruno Després	*LJLL, Sorbonne University (Paris)*
	Sergio Amat Plata	*DMAE, UPCT (Cartagena)*
Examinatrice:	Francesca Rapetti	*Lab. Math. J. A. Dieudonné (Nice)*
Directeurs:	Eric Serre	*CNRS-M2P2 (Marseille)*
	Frédéric Schwander	*M2P2, ECM (Marseille)*
	Jacques Liandrat	*I2M, ECM (Marseille)*
Membres invités:	Patrick Tamain	*IRFM-CEA (St. Paul-lez-Durance)*
	Giorgio Giorgiani	*M2P2 (Marseille)*

Abstract

Highly anisotropic elliptic problems occur in many physical models that need to be solved numerically. In the problems investigated in this thesis, a direction of dominant diffusion exists (called here parallel direction), along which the diffusivity is several orders of magnitude larger than in the perpendicular direction. In this case, standard finite-difference methods are generally not designed to provide an optimal discretization and may lead to the perpendicular diffusion being artificially supplemented by a potentially large contribution stemming from errors in approximating parallel diffusion.

This thesis focuses on three main axes to suitably solve anisotropic elliptic equations: an aligned, conservative finite-difference scheme to discretize the Laplacian operator, a reformulated Helmholtz equation to avoid spurious numerical diffusion, and a solver based on multigrid methods as a preconditioner of GMRES routine. Although the scope of this thesis is the application on plasma edge physics, results are relevant to any highly anisotropic model flow in bounded domains.

In Chapter 1, a short introduction to magnetically confined fusion is presented identifying the numerical problems raised by solving fluid equations, in particular in the Scrape-Off Layer region. The numerical problem which is dealt with is an anisotropic elliptic problem where diffusivity is 5 to 8 orders of magnitude larger in the parallel direction. This large parallel diffusivity results in long wavelengths in the parallel direction, a central characteristic to the understanding of methods discussed in this thesis. In Chapter 2, a bibliographic introduction to numerical methods dedicated to the solution of anisotropic elliptic equations is presented, with a focus on finite-difference methods. Aligned methods, and their potential to compute solutions with accuracy comparable to standard methods with much lower number of mesh points are presented.

In Chapter 3 we propose an original aligned discretization scheme using non-aligned Cartesian grids. Based on the Support Operator Method, the self-adjointness of the parallel diffusion operator is maintained at the discrete level. Compared with existing methods, the present formulation further guarantees the conservativity of the fluxes in both parallel and perpendicular directions. For bounded domains, a discretization of boundary conditions is presented ensuring comparable accuracy of the solution. Numerical tests based on manufactured solutions show that the method provides accurate and stable numerical approximations in both periodic and bounded domains with a drastically reduced number of degrees of freedom with respect to non-aligned approaches.

A reformulation of the Helmholtz equation is presented in Chapter 4 to limit spurious numerical diffusion. The method is based on splitting of the original problem into two distinct problems for the aligned and the non-aligned parts of the solution. These two contributions are separated by filtering methods which are evaluated. Tests cases show

this reformulation eliminates spurious perpendicular diffusion, with larger impact on accuracy with higher parallel diffusivities.

Finally, with the aim of solving elliptic anisotropic equations for large systems efficiently, a geometric multigrid algorithm is proposed in Chapter 5 in bounded domains. The algorithm scales adequately with the number of degrees of freedom, and shows a clear advantage upon standard iterative methods when the parallel diffusivity is very large. This algorithm is later posed as preconditioner of a GMRES solver, and its efficiency is compared with that of direct solvers solving elliptic equations under any boundary conditions.

The thesis is concluded by a critical analysis of the numerical aspects of aligned discretizations investigated. Special attention is given to the application of the investigated schemes in 3D plasma turbulence codes, such as the TOKAM3X developed by CEA.

Résumé

Les problèmes elliptiques hautement anisotropes se présentent dans de nombreux modèles physiques qui doivent être résolus numériquement. Une direction de diffusion dominante est alors introduite (appelée ici direction parallèle) le long de laquelle le coefficient de diffusion est plusieurs ordres de grandeur plus grand que dans la direction perpendiculaire. Dans ce cas, les méthodes aux différences finies standard ne sont pas conçues pour fournir une discrétisation optimale et peuvent conduire à une diffusion perpendiculaire artificielle potentiellement importante, résultant en un erreur significative dans l'approximation de la diffusion parallèle.

Cette thèse se concentre sur trois axes principaux pour résoudre les équations elliptiques anisotropes de manière appropriée : un schéma aligné et conservatif de différences finies pour discrétiser l'opérateur Laplacien, une reformulation de l'équation de Helmholtz pour réduire la diffusion numérique, et un solveur basé sur les méthodes multi-grille comme préconditionneur d'un solveur GMRES. Les deux premiers chapitres sont consacrés à la présentation du cadre de cette thèse.

Au chapitre 1, une brève introduction à la fusion par confinement magnétique est présentée, identifiant les problèmes numériques soulevés par la résolution des équations fluides, en particulier dans la région proche au bord (Scrape-Off-Layer). Le problème numérique que nous allons traiter est essentiellement un problème elliptique anisotrope où la diffusion est de 5 à 8 ordres de grandeur plus grande dans la direction parallèle que dans la direction perpendiculaire.

Dans le chapitre 2, une introduction bibliographique aux méthodes numériques résolvant les équations elliptiques anisotropes est présentée, avec un accent sur les méthodes aux différences finies.

Dans le chapitre 3, un schéma de discrétisation aligné est proposé en utilisant des grilles cartésiennes non alignées. Selon la méthode Support Operator Method (SOM), la propriété que l'opérateur de diffusion parallèle est auto-adjoint est maintenue au niveau discret. Par rapport aux méthodes existantes, la formulation actuelle garantit la conservation des flux dans les directions parallèles et perpendiculaires. De plus, dans les domaines bornés, une discrétisation des conditions aux limites est présentée afin d'assurer une précision comparable de la solution. Des tests numériques basés sur des solutions manufacturées montrent que la méthode est capable de fournir des approximations numériques précises et stables dans des domaines périodiques ou bornés avec un nombre considérablement réduit de degrés de liberté par rapport aux autres approches non alignées.

Une reformulation de l'équation de Helmholtz est présentée au chapitre 4 pour limiter la diffusion numérique liée à la discrétisation du Laplacien pour les valeurs élevées de

diffusion parallèle. La méthode est basée sur la séparation de la solution endeux contributions (alignée et non alignée) par rapport à l'opérateur de diffusion parallèle, grâce à des méthodes de filtrage. Les cas de test montrent que cette reformulation de l'équation de Helmholtz élimine la diffusion numérique perpendiculaire, avec un impact d'autant plus accrue que les valeurs de diffusivité parallèle sont élevées.

Afin de résoudre efficacement les équations anisotropes elliptiques pour les grands systèmes d'équations, un solveur itératif basé sur des algorithmes multi-grilles géométriques est proposé au chapitre 5. Cet algorithme est plus tard posé comme préconditionneur d'un solveur GMRES, exhibant une réduction drastique du temps et de la mémoire requise par rapport à des solveurs directs résolvant les équations Helmholtz et Poisson, et ce pour différents types de conditions aux limites.

La thèse est conclue par une analyse critique des aspects numériques des discrétisations alignées étudiées. Une attention particulière est accordée à l'application des méthodes étudiées dans les codes de turbulence plasma 3D, tels que TOKAM3X développé par le CEA.

Remerciements

D'abord, j'aimerais remercier les membres du jury: M. B. Després, M. Amat Plata, et Mme. Rapetti pour avoir montré de l'intérêt pour mon travail, et être venu de si loin pour ma soutenance de thèse. Merci!

J'aimerais bien sûr remercier le travail (parfois très dur) de mes directeurs de thèse. Eric Serre, merci pour votre sérénité au moment de faire des corrections et de m'avoir guidé sur le bon chemin de l'expression (S_a ou T_a, c'est toujours la question). Frédéric Schwander: merci de votre patience au moment de m'expliquer 50 fois la même chose, c'était un honneur pour moi d'avoir travaillé avec un professeur de votre niveau de connaissance.

Parallèlement j'ai aussi reçu la base numérique de mon travail au travers de Jacques Liandrat. Merci aussi parce que je suis là grâce à une simple phrase qui a vraiment changé ma vie: *Tu es peut être intéressé pour faire le Master Fusion, tu peux parler avec Guido Ciraolo et postuler...*

Merci aussi à Giorgio Giorgiani, pour m'avoir donné ton avis toujours très direct (avec les avantages donnés par sa connaissance du langage Espagnol le plus pur).

Bien sûr, je voudrais remercier les membres du CEA avec lesquels j'ai eu l'honneur de travailler et de discuter comme Patrick Tamain et Guido Ciraolo.

Le déroulement de cette thèse s'est fait dans un environnement le plus approprié: le laboratoire M2P2. Merci au directeur du laboratoire, M. Sagaut, à tous les membres permanents, l'administration (Sadia, Sophie), Elena et Michel pour votre disponibilité et accueil.

Pendant toutes ces années, j'ai eu l'honneur de travailler avec des collègues qui marqueront le futur de cette profession *sans aucun doute*. Davide *Galaxy*, Matteo, Rachel (I miss whistling *My boy lollipop*), Eddy (*Edouard*) Constant, Sylvain (et aussi Rena), Tovarish Alexandrovich, la reine Roua, Dr. Jinming Lyu, Carlos, Tayyab, *Momo!*, Benjamin, Mme PhD Guiza, Gabriel, Raffaele (oh la la...), Thomas, Nicola, (Mr. President) N. Franghief, et Sylvia.

Je sors d'ici avec certaines amitiés pour tout la vie: Marianna Peponita (et sa mère), Eunok Yim (prononcé *iime*) et Elisabetta Chaschera (de Maceratta).

Si je suis là c'est aussi grâce à un grand collègue du master qui m'a beaucoup aidé pour m'adapter à la vie en France: merci Sylvain Hérault pour tout. Merci aussi aux collègues du Master Fusion (mention spétiale à Aissa *Gran* Kali et Javier).

Merci aussi à mes principaux supporters espagnols: Merci Javi, Consue et Rubia. Merci à mes amis de tout ma vie: Rob, Juen, Mario et Antonio.

Merci aussi aux professeurs qui m'ont guidé vers le chemin de la science: Merci M. Lopez Espin, M. Zamora Barrancos, M. Amat Plata, Mme Toral Noguera.

En ces moments je n'oublie pas mes grand-parents, Rosario, Eduardo et Antonio: leurs vies sont toujours une importante inspiration pour moi.

A mes parents: mes sœurs et moi nous sommes bien conscients de vos efforts pour nous donner le futur qu'on a aujourd'hui. Merci Tatica, Beatrícheri et aussi à un membre attaché à ma famille: mon presque frère *El Papi*. Merci Yuki de t'intercaler entre moi et l'écran.

Mais surtout à toi. Merci Angela, la pièce fondamentale de ma vie.

Contents

Chapter 1

A brief introduction on magnetic fusion

This work concerns the development of advanced numerical schemes for the simulation of plasmas for magnetic fusion. This chapter briefly depicts the main physical principles of fusion, and provides the set of fluid equations implemented in TOKAM3X [TBC+16], the code currently developed by the team to simulate plasma flows at the edge of the tokamak chamber. One of the specificity of this configuration is the strong anisotropy introduced by the magnetic field between its parallel and transverse direction in the plasma flow. Mathematically, this leads to strongly anisotropic differential operators into the equations. This work is motivated by the need to design more efficient numerical schemes allowing one to satisfy resolution and accuracy requirements to perform reliable simulations of turbulent plasma in realistic magnetic configurations. The reader is referred to the works of Ref. [Wes97, Fre07, GR07, Dav01] for more information.

1.1 The magnetic confinement fusion

1.1.1 Conditions for fusion

Nuclear fusion is a reaction in which two or more atomic nuclei are combined to form one or more different atomic nuclei and subatomic particles (neutrons or protons). The difference in mass between the reactants and products is manifested as either the release or absorption of energy. During ITER operation, the fusion reaction will involve two hydrogen isotopes, deuterium (^2D) and tritium (^3T) leading to:

$$^2D + {}^3T \longrightarrow {}^4He + {}^1n \tag{1.1}$$

with a mass defect of $3,1 \times 10^{-29}$ kg leading an energy production of $E = 17,6$ MeV according to the famous Einstein's equation, $\Delta E = \Delta m c^2$. This amount of produced

energy corresponds to the sum of the kinetic energy of both products, i.e. $3,5$ MeV (^4He) and $14,1$ MeV (^1n).

Critical conditions to reach fusion (also known as Ignition condition) of a D-T plasma have been identified by Lawson [Law57]. The *Lawson criterion* establishes a condition on the product of the plasma temperature T, the plasma density n and a characteristic time for the plasma called confinement time τ_e that must satisfy the inequality:

$$n\,T\,\tau_e \geq 3 \times 10^{21}\text{keV} \cdot \text{s} \cdot \text{m}^{-3} \tag{1.2}$$

As the density is constrained in magnetically confinement fusion [GTW+88], critical conditions to reach fusion require high temperatures and large confinement time. In order to give an idea of the order of magnitudes of the different physical quantities, for a plasma at the pressure $n\,k_B\,T \approx 1,6{\cdot}10^5 Pa$, the criterion 1.2 is satisfied for $n = 10^{20}\text{m}^{-3}$, $k_B\,T \simeq 10^8 K$ and $\tau_E > 3$s.

1.1.2 Single particle motion

The motion of a single charged particle with velocity $\mathbf{v}$ in a magnetic field $\mathbf{B}$ is driven by the Lorentz force, $\mathbf{F} = q(\mathbf{v} \times \mathbf{B})$ (q being the electric charge of the particle). This force acts perpendicular to the direction of motion, causing the particle to gyrate, or move in a circle as sketched on Fig. 1.1.

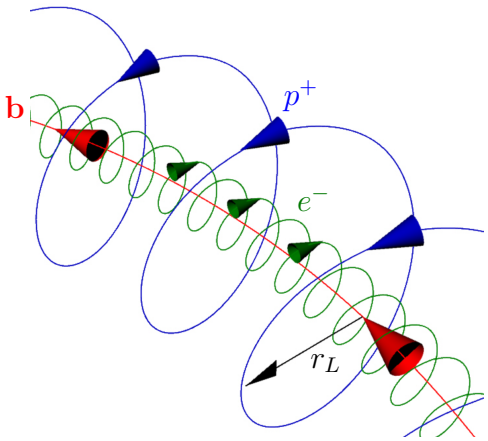


Figure 1.1: Chart of ion (blue trace) and electron (green trace) trajectory along a magnetic field line (red trace). Note that the ion Larmor radius ρ_L is larger than the electron one due to the mass difference ($m_i \gg m_e$, Eq. 1.3). $\mathbf{b} = \mathbf{B}/\|B\|$ is the unit vector along the magnetic field line. The drift velocity is counter-directional (since the electric charge $e_i = -e_e$)

In the case where the magnetic field is uniform in space and time and in absence of electric field, the radius of this circle, called gyroradius or Larmor radius can be expressed as:

$$\rho_L = \frac{m|v_{b\perp}|}{|q|B} \tag{1.3}$$

where m is the mass of the particle and $|v_{b\perp}|$ is the magnitude of the component of the velocity in the plane perpendicular to the magnetic field line. This is a fundamental principle of confinement exploited in magnetic fusion devices: the particle motion in the perpendicular direction is bounded around the field-line.

1.1.3 The tokamak configuration

Different principles of magnetic confinement devices have been tested in the past. The tokamak (Fig. 1.2) exploits the feature of a toroidal geometry, which allows to bend a magnetic field line and close it on itself to constrain charged particles to spend enough time along the magnetic field-lines to hope obtaining a sufficient confinement time τ_E, Eq. 1.2. However, this description corresponds to an ideal situation. Indeed in presence of electric field and curved magnetic field lines, the guiding center of the charged particles trajectory, which is the center of the cyclotron gyration described on Fig. 1.1, drifts. This drift motion is characterized by the so-called drift velocities, which can be expressed analytically depending on the physical mechanism at play [GB14]:

- The electric drift velocity (identical for ion and electron):

$$\mathbf{u}_{E\times B} = \frac{\mathbf{E}\times\mathbf{B}}{B^2} \tag{1.4}$$

  It is generated by the electric force, independently of the mass of the particle and of the particle charge, thus not causing any net current.

- The curvature drift velocity :

$$\mathbf{u}_{\triangledown B} = \frac{mv_{b\|}^2}{qB^2}\mathbf{B}\times(\mathbf{b}\cdot\nabla)\mathbf{b} \tag{1.5}$$

  It is generated by the magnetic field line curvature. This drift velocity is opposite for ion and electron due to the charge q, and it is vertical in the case of a toroidal field.

Remark: there is another drift velocity called polarization velocity but being second-order in a normalized gyroradius expansion it is usually neglected [TGT+09].

To balance the vertical motion induced by the velocity drifts, it is thus necessary to add in the tokamak another component of the magnetic field in the poloidal direction able to limit these losses (Fig. 1.2). This component is mainly produced by the electric current flowing in the plasma. The resulting magnetic field lines have a helicoidal shape (red lines in Fig. 1.2). Each magnetic line is defined in a closed surface, defined by R (large radius of the torus), and a (the small radius of the torus). On a magnetic surface, at a given radial position r, the position is given by a toroidal (φ) and a poloidal (θ) angle.

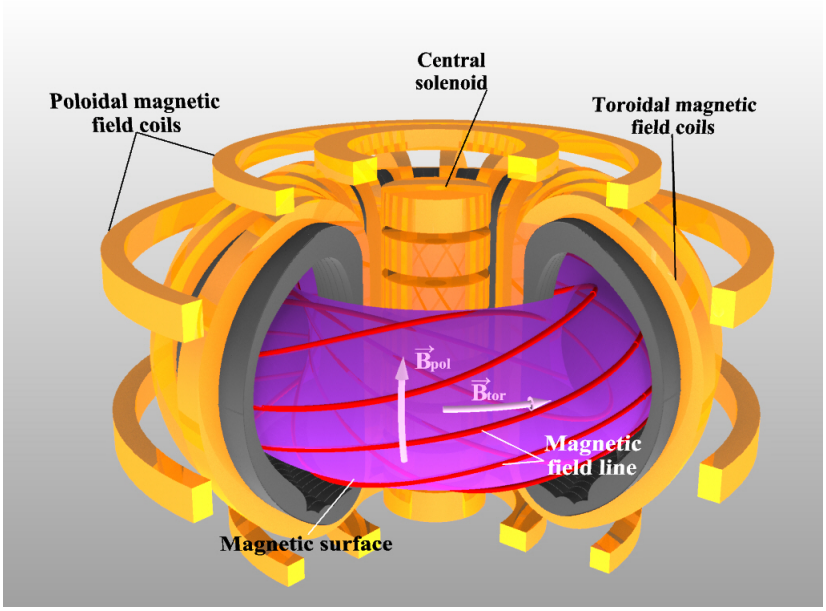


Figure 1.2: Sketch of the tokamak magnetic configuration. At a given r, magnetic field lines (red lines) roll up around the magnetic surface (purple).

The relative amplitude of the toroidal and poloidal components of the magnetic field determines the local inclination of the field line with respect to the toroidal direction, which is called pitch angle. This parameter has to be taken into account in the design of the numerical scheme. Following the direction along the field-line, we can quantify the number of toroidal turns completed before returning to the initial poloidal position. This quantity is called safety factor and it is defined as:

$$q = \frac{\Delta\varphi}{2\pi}, \tag{1.6}$$

where $\Delta\varphi$ is the variation angle in the toroidal direction. In a simplified geometry with a torus of circular cross-section, q simplifies to:

$$q = \frac{B_\varphi r}{B_\theta R} \tag{1.7}$$

Considering the poloidal magnetic field to be negligibly small when compared to the toroidal component, one can provide an approximation of the total *parallel length* $L_{b\|}$ of a field-line as:

$$L_{b\|} \sim 2\pi qR \tag{1.8}$$

This estimate gives the longest spatial scale to discretize in the simulation.

Ideally, the confined plasma should be completely isolated from the walls of the vacuum chamber. However, it is technically impossible to build a magnetic field tangential at all the points to a certain surface and so magnetic field-lines unavoidably intercept a solid component at some point. The motion of particles along the field lines therefore leads to non-zero particle and heat fluxes on the wall components, which must be bounded in order to avoid damage to the materials. The simplest technical solution to control these fluxes to the wall components is called the *limiter* configuration. It corresponds to the insertion of a solid component into the tokamak chamber to be intercepted by the magnetic filed-lines. Usually, this limiter is toroidal and extends uniformly into the toroidal direction. Field-lines that intercept the solid components are commonly called (in a somehow misleading way) *open field-lines*, and the plasma region defined by these lines *Scrape-Off Layer* (SOL). This region is separated from the closed field-lines by a magnetic surface named *Last Closed Flux Surface*.

In more recent tokamaks, a more efficient configuration to keep the plasma core away from plasma wall interactions is the *divertor* configuration. A purely toroidal field-line is generated, whose projection on the poloidal plane is called *X-point*. At this point, the poloidal magnetic field is null and the magnetic field-line is theoretically of infinite length, as well as the safety factor q. A schematic representation of the divertor configuration is given in Fig. 1.3.

The plasma edge - The team research activity focuses on the plasma edge region, which encompasses the open field lines (SOL) and the outer part of the closed field region on both sides of the Last Closed Flux Surface or *separatrix*, Fig. 1.3. The dynamics of the plasma in this region plays a crucial role in the tokamak exhaust system, in plasma refuelling, and in the dynamic of impurities [GB14, SM90, Sta00].

This plasma-wall interaction also involves complex atomic processes which lead to multi-physics problems and thus to challenging numerical issues to properly model the plasma flow in this region.

1.1.4 Transverse transport in tokamak plasmas

In order to understand the global confinement in a tokamak, one has to evaluate the collective characteristics of the plasma, including the interactions among particles. One is then interested in understanding how the tokamak global system can maintain high particle density and temperature at its centre, whilst maintaining reasonable characteristics

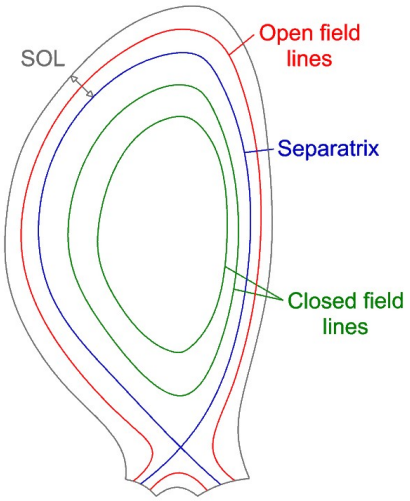


Figure 1.3: Sketch of the magnetic surfaces at the plasma edge. The *Scrape Off Layer* (SOL) corresponds to open field lines from the *separatrix* (last closed magnetic line) to the wall.

at its peripheral region to deal with the presence of solid components.

High temperatures around $1.5 - 3 \times 10^8$K are required to reach fusion conditions [org19]. The high kinetic energy of particles combined with a low particle density render collisions inefficient, and the mean free path between collisions in the parallel direction can be of the order of the parallel length $L_{b\|}$ (several meters, m). On the other hand, the confinement by the magnetic field restricts the mean free path in the perpendicular direction to the Larmor radius. However, physical phenomena can lead to the migration of particles or energy from the centre to the SOL region across magnetic surfaces. Here, we briefly list these *transverse transport* processes, and for more details, the reader is referred to Ref. [Gar01, Tam07].

- The so-called *classical* transport is caused by Coulomb collisions between charged particles in a uniform magnetic field. It gives rise to a diffusion that can be characterized by a collision frequency ν_{coll} ($\sim nT^{-3/2}$) to displace a particle at the distance ρ_L: $D_{coll} \sim \nu_{coll}\rho_L^2$. Typical values for deuterium ions range from 10^{-4} to $10^{-2}m^2s^{-1}$.

- The so-called *neoclassical* transport is a more advanced evaluation taking into account the real shape of magnetic field lines, and especially the dependence on $1/R$ of the magnetic field amplitude in the toroidal direction. It leads to some adjustments in the definition of the diffusion coefficient defined above taking into account different regimes for the particles. Typical values of diffusion coefficient are typically between 10 and 100 times larger than for the classical transport.

- The so-called *anomalous* transport is related to turbulence, Fig. 1.4. It is defined

as anomalous because experiments show a much higher thermal and mass loss than theoretically predicted by neoclassical transport. Typical values of diffusion coefficients measured in experiment are of the order of $1m^2s^{-1}$.

1.2 The TOKAM3X fluid model

A proper understanding of the edge would require *full-f gyrokinetics* simulations based on the distribution function. Pioneering full-f gyrokinetic simulations of the edge start appearing in the fusion community, addressing physical phenomena of fundamental interest for fusion operation like transport barrier formation [CKG$^+$06, CKT$^+$17]. However, despite the exponential growth of computer speed along with significant improvements in computer technology, they remain extremely costly from the computational point of view. It is particularly true in the near-wall region where particle recirculation requires addressing the electron and ion dynamics on the same footing, and in a magnetic topology that is much more complex than in the core. As a consequence, though approximate, the fluid approach remains a standard one near the wall where the temperature is lower and the collisional mean free path significantly smaller than in the core.

Three-dimensional fluid conservation equations are obtained for electrons and ions using simplified closures developed by Braginskii [Bra65]. The model presented below is the one implemented in the isothermal version of the TOKAM3X code developed for many years by the laboratory in close collaboration with CEA, see in Ref. [Tam07, TBC$^+$16, Col15, GTB$^+$17, TGT$^+$10]. Under some hypotheses and ordering detailed below, four equations are derived for four unknown dimensionless fields: the electronic density N, the ionic parallel momentum Γ, the electrostatic potential Φ and the parallel current $j_{b\parallel}$ which defines the parallel advection velocity for electrons.

1.2.1 Hypotheses and ordering

- *The quasi-neutrality assumption:* $n_e \approx Zn_i$, the smallest turbulent scales (of the order of ten ion Larmor radius ρ_L) being much larger than the Debye length λ_D above which electric charges are screened, $\rho_L \gg \lambda_D$. The very near wall sheath region is not directly described and appears thanks to Bohm boundary conditions (see in [Sta00] and thereafter for details). Here, the assumption is restricted to hydrogen ions, $Z = 1$.

- *The electron inertia is negelected*: $m_e/m_i \simeq O(10^{-3}) \Longrightarrow m_e \ll m_i$,

- *The drift ordering* (see for example in Ref. [TGT$^+$09]). It is based on assuming that the characteristic plasma frequencies ω are slow compared to the ion cyclotronic frequency ω_{ci}, $\omega \ll \omega_{ci}$. This frequency assumption leads to a strong scale separation between the Larmor radius ρ_L ($\simeq 1mm$) and a characteristic length of

turbulence structures ($\simeq 0.1 - 10cm$). Within the drift ordering, it is useful to split the analysis of the dynamics into the parallel and perpendicular directions to the magnetic field, by decomposing the velocity for the ions and the electrons,

$$\mathbf{u}^i = u^i_{b\|}\boldsymbol{b} + \mathbf{u}^i_{b\perp} \quad and \quad \mathbf{u}^e = u^e_{b\|}\boldsymbol{b} + \mathbf{u}^e_{b\perp} \tag{1.9}$$

where the perpendicular components of the velocity are described analytically in terms of drifts:

$$\mathbf{u}^i_{b\perp} = \mathbf{u}_{E\times B} + \mathbf{u}^i_{\nabla B} + \quad and \quad \mathbf{u}^e_{b\perp} = \mathbf{u}_{E\times B} + \mathbf{u}^e_{\nabla B} \tag{1.10}$$

- These velocity components leads to a total current expression $\vec{j}$:

$$\boldsymbol{j} = j_{b\|}\boldsymbol{b} + Ne(\mathbf{u}^i_{\nabla B} - \mathbf{u}^e_{\nabla B}), \tag{1.11}$$

  which is the addition of the parallel and diamagnetic currents.

- *The model is isothermal.* T_i and T_e are given by a steady arbitrary temperature spatial distribution for ions and electrons.

- *The plasma is assumed to be electrostatic.* The plasma magnetic pressure is assumed to be much higher than the kinetic pressure. Then, the effect of the magnetic fluctuations on transport are negligible and only the fluctuations of the electric potential are taken into account.

1.2.2 Fluid equations

The hypotheses enunciated above lead to a conservation equation for the electronic density N (for simplicity with respect to ionic density equation), for the ionic parallel momentum Γ (obtained by summation of the equations for the ions and the electrons) and for the vorticity W, which replaces the charge balance equation ($\nabla.\mathbf{j} = 0$), see in Ref. [TBC+16]:

$$\begin{cases} \partial_t N + \nabla\cdot\Big(N\mathbf{u}^e\Big) = S_N + \nabla\cdot(D_N\nabla_{b\perp}N) \\ \partial_t\Gamma + \nabla\cdot\Big(\Gamma\mathbf{u}^i\Big) = -\nabla_{b\|}P + \nabla\cdot(D_\Gamma\nabla_{b\perp}\Gamma) \\ \partial_t W + \nabla\cdot(W\mathbf{u}^i) = \nabla\cdot\Big(N(\mathbf{u}^i_{\nabla B} - \mathbf{u}^e_{\nabla B}) + j_{b\|}\mathbf{b}\Big) + \nabla\cdot(D_W\nabla_{b\perp}W) \\ with \\ j_{b\|} = -\frac{1}{\eta_{b\|}}\nabla_{b\|}\phi + \frac{1}{N\eta_{b\|}}\nabla_{b\|}N \\ W = \nabla\cdot(\frac{1}{B^2}(\nabla_{b\perp}\phi + \frac{1}{N}\nabla_{b\perp}N)) \end{cases} \tag{1.12}$$

where $\eta_\parallel$ is the normalized parallel collisional resistivity of the plasma, S_N a volumetric source term included to drive the particle flux, and $\nabla_{b\parallel}$ and $\nabla_{b\perp}$ are the parallel and perpendicular gradients respectively to $\mathbf{b}$. The driving term in the momentum equation is the dimensionless parallel static pressure gradient $\nabla_{b\parallel} P$.

In all equations above, effective diffusion terms account for collisional transport and roughly model the effect of turbulent small scales (smaller than the grid spacing) in the cross-field direction, see Fig. 1.4. Their parallel component has been neglected compared with parallel convection. $D_{N,\Gamma,W}$ are arbitrary constants, usually smaller than 1.

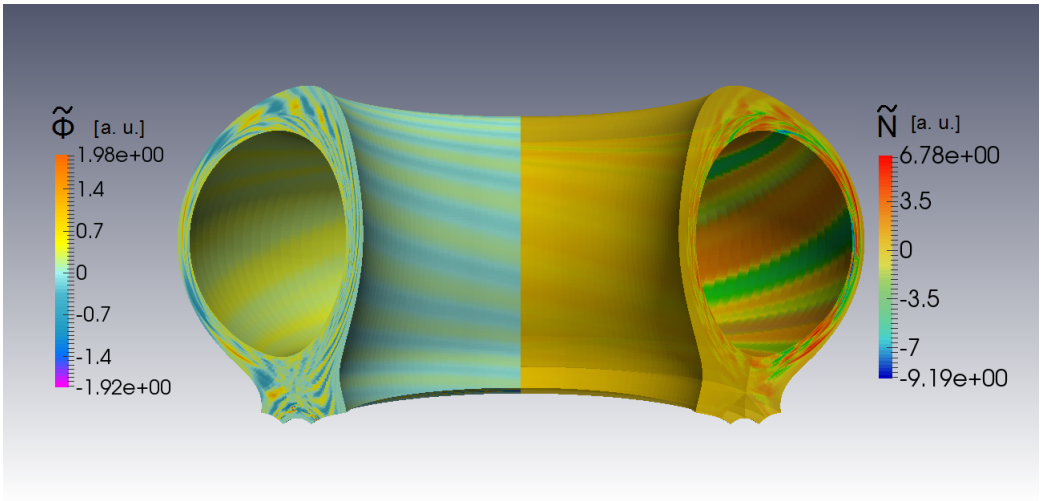


Figure 1.4: Example of turbulent structures in a TOKAM3X simulation in a divertor configuration [TBC+16, TGT+09]. Highly anisotropic transport and diffusion generate a flow characterized by elongated structures along the magnetic field lines, in combination with rapid spatial variations in the perpendicular direction due generated by turbulence. Fluctuations of electric potential (left) and density (right).

Boundary conditions The following boundary conditions are associated to the set of conservation equations in the radial and parallel direction.

At the radial boundary conditions, both at the core and at the external wall,

$$\partial_{b\perp}(\cdot) = 0 \tag{1.13}$$

In the parallel direction at the targets where the field lines intercept the wall:

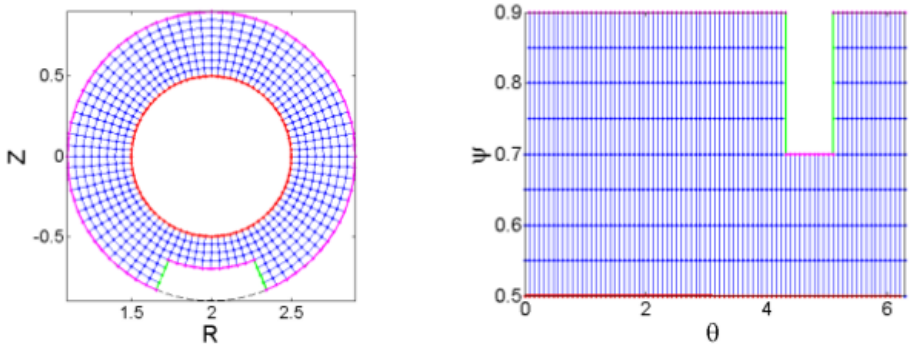
$$\begin{cases} |\Gamma| \geq N \\ \nabla_{b\|}\phi = \pm\eta_{b\|}N(\Lambda - \phi) + \frac{\nabla_{b\|}N}{N} \\ \partial_\theta^2 N = 0 \\ \partial_\theta W = 0 \end{cases} \tag{1.14}$$

where Λ is the sheath floating potential. This later corresponds to the usual Bohm boundary conditions [SM90] and models the physics of the sheath located next to the limiter wall.

1.2.3 Numerical schemes

As mentioned above, the flow in the parallel direction corresponds to a compressible gas flow, whereas in the perpendicular direction it corresponds to an quasi-incompressible flow as a result of the strong magnetic field, dominated by turbulent processes. In addition, the magnetic topology in tokamak edge is complex and makes the flow strongly anisotropic. Therefore, the conservation equations presented above and governing the edge/SOL plasma require specific algorithms that usually split the discretization of the parallel and perpendicular directions.

In order to limit numerical diffusion (see the next chapters for more details) the equations above are discretized over a structured magnetic flux-surface aligned grid where the first direction of the mesh is along iso-ψ lines in the poloidal plane (see 2 examples in Figs. 1.5 and 1.6). Here, ψ denotes the orthogonal coordinate to the magnetic surfaces, which corresponds to the radial r coordinate in circular cross-section, see [TGT+09].



(a) (b)

Figure 1.5: Examples of mesh in limited circular cross-section. (a) Mesh distribution in the physical (R, Z)-plane (left) and in the (ψ, θ)-plane (right). The limiter is located at the bottom of the machine.

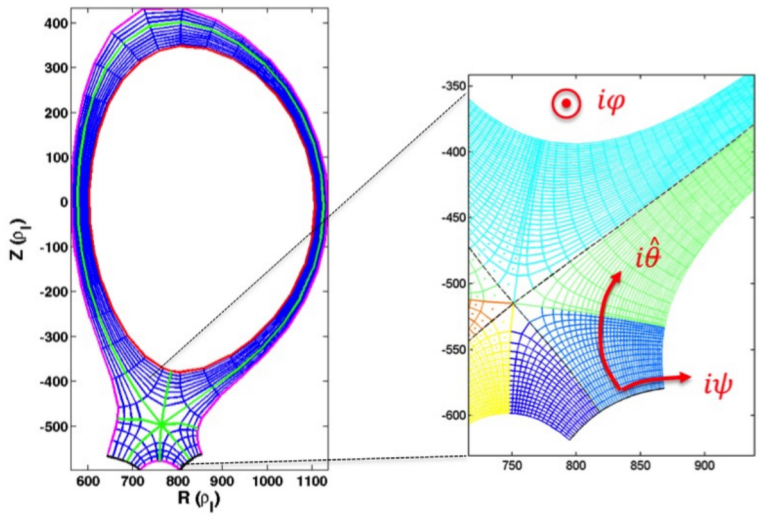


Figure 1.6: Example of mesh in diverted poloidal cross-section. Mesh distribution in the physical space of a diverted geometry emphasizing the domain decomposition (green lines) and the grid distribution around the X-point.

The spatial discretization is based on a second-order conservative finite-differences scheme associated to a 3rd-order WENO reconstruction for the advection terms to deal with both shocks and complicated structures of the solution [LOC94]. The time evolution is based on a first-order operator splitting. Implicit and explicit terms are the following:

- Advection and source terms are mainly non-linear. Their dynamics is on ionic time scale which allows an explicit advancement.
- The parallel current terms express the evolution of the plasma electric potential. They are advanced using a fully implicit 3D solver in order to capture the associated fast dynamics without considerably constraining the time step.
- The perpendicular diffusion terms are advanced implicitly in order to allow large diffusion coefficient, running the code in transport mode (i.e. no turbulent small scales).

For all these terms, the advancement of parallel current terms in the vorticity equation is the main numerical issue. The time evolution of the plasma potential writes as :

$$(\mathcal{L}^{b\perp} + \ \delta t\ \mathcal{L}^{b\parallel})\phi^{\star\star} = W^{\star} - \mathcal{L}^{b\perp}N^{\star} + \ \delta t\ \mathcal{L}^{b\parallel} lnN^{\star} \tag{1.15}$$

where $\mathcal{L}^{b\perp,b\parallel}$ are spatial differential operators defined as $\mathcal{L}^{b\perp} = \nabla \cdot (\frac{1}{B^2}\nabla_{b\perp}\cdot)$ and $\mathcal{L}^{b\parallel} = \nabla \cdot (\frac{1}{\eta_{b\parallel}} b\nabla_{b\parallel}\cdot)$

Eq. 1.15 associated to the set of complex boundary conditions detailed above, perfectly illustrates the kind of numerical problem that we want to address in this PhD work. The very small value of the parallel resistivity in tokamak plasma ($\eta_{b\|} \approx 10^{-5} - 10^{-8}$ (normalized values)) leads to a strongly anisotropic and very badly conditioned differential operator. The discretization of such an operator requires a well-adapted numerical scheme to limit numerical diffusion without doing an unaffordable effort on the number of degree of freedom that will make the computation too costly.

Chapter 2

Numerical discretizations of an anisotropic diffusion problem

The present chapter aims to present a general overview on all numerical methods devoted to the resolution of anisotropic elliptic equations, with a specific focus on finite-difference schemes.

2.1 Introduction

Elliptic partial differential systems are ubiquitous in physical models and numerical simulations. They occur in fluid models used in mechanics, geophysics, plasma physics, but also in other fields of research as in microelectronics, optics or image processing, the list being not exhaustive.

No universal method exists that provides efficiency and accuracy in the resolution of such a problem, but a variety of methods, often depending on the spatial discretization scheme.

For finite-element, Le Poitier and Hai Ong [LPHO12] introduced the conservative finite element cell-centered method. Its interest is to lead to a symmetric positive definite matrix at the discrete level and a reduced stencil. In their work, Baliga *et al.* [BP83] presented an hybrid CV-FE methodology, combining the conservative properties of cell-centred control volume approach and finite element meshes. Although this technique has been applied in several fields [PT99, FT96] with satisfactory results, it becomes costly in terms of computational time calculation when using fine meshes, especially in high anisotropic ratios or strongly orthotropic rate [JT03]. Jayantha *et al.* [JT05] proposed a finite volume discretization method on unstructured meshes. Since high anisotropy has been obtained by hybrid control-volume / finite-element methods to evaluate the fluxes through the cells, very fine meshes must be considered to obtain satisfactory precision. The results show a better accuracy in 2D cases with high anisotropy rate ($K_{xx}/K_{yy} \approx 10^3$-

10^4). Finally, the discontinuous Galerkin (DG) methods have the potential to lead to a stable and high-order discretization for elliptic problems [CS97, BS92, PP08, FKX13]. More recently, hybridized versions of Discontinuous Galerkin (HDG) [CGL09, NPC09], allowed to reduce the coupled degrees of freedom of DG methods, and allowed the use of adapted methods in different parts of the discrete domain. Due to this, HDG methods show strong stability properties, wide adaptability to parallelized solvers, and superconvergence properties [CDG08, NPC11, CC12, CC14]. In the team, Giorgiani *et al.* [GBC+18] validated results on 2D reduced models derived from fluid transport equations for the plasma edge, presenting a highly reduced numerical error thanks to the use of high order numerical schemes. The use of such unstructured meshes not aligned with the magnetic field allows an accurate modelling of realistic plasma chambers.

For finite-volume, Herbin et Hubert [HH08] presented a detailed benchmark comparing 25 numerical discretizations found in the literature in 2D domains, and they concluded into a generalized homogeneity in the results. This work has been lately carried out in 3D by Hubert *et al.* [EHH+11]. We can mention the work of Le Poitier [LP05] who introduced a FVM using triangular cells, and in which gradients are calculated by nonlinear schemes and where the minimum-maximum principles are satisfied. The method is shown to be robust and efficient compared with methods for which the minimum-maximum principles are not satisfied. Introduced by Aavastsmark *et al.* [ABBM94, ABBM96] for multiphase and anisotropic petrol reservoirs simulations, MPFA leads to a conservative finite-volume discretization of flow equations for general non-orthogonal grids, as well as for anisotropic orientation of the permeability tensor. In this method, the conservative flux definition is built considering an interaction region between adjacent cells, where a set of transmission coefficients are established in function of the cells and K orientation. Since MPFA has been developed over years varying the flux definition [Aav07], the method is conditioned to accomplish monotonicity in the whole domain and in the grid construction. In 3D applications, however, convergence proofs do not exist. The method is robust in terms of diffusion tensor discontinuity, but the symmetry of the discrete diffusion operator is not assured, and the accuracy decays when the anisotropy level becomes large. Maire *et al.* [MB11, MB12] introduced the *Cell-Centered Lagrangian Diffusion* (CCLAD) method, which leads to a second order accurate FVM. CCLAD is assembled considering cell-centred unknowns and a local stencil, introducing in the cell interface two half-edge normal fluxes and temperatures, the fluxes being approximated by sub-cell variation (CCLADS) or a finite difference approximation (CCLADNS). The method has been later adapted to 3D by Jacq *et al.* [JMA13] obtaining a symmetric positive definite matrix. In their work, Hermeline [Her00], introduced the Discrete Duality Finite Volume (DDFV), which independently of the mesh regularity, leads to positive definite matrix ensuring a second-order accuracy. Applications of DDVF with discontinuous diffusion tensors have been introduced by Hermeline [Her03]. More recently, Gander *et al.* [GHHK18] introduced DDFV for anisotropic diffusion, implementing Optimized Schwartz methods to obtain effective transmission conditions at

the cell interfaces. Finally, a comparative benchmark on FVs was presented by Droniou [Dro14], focusing on the comparison in coercivity property, which ensures the method's convergence, and the minimum-maximum property, crucial for high anisotropy. On another side, a generalized comparison on locally conservative methods was made by Klausen and Russell [KR05] for reservoir applications. All those comparisons concluded that a method applicable to any circumstance does not exist, finding convergence slopes related to the method order.

To conclude this brief overview, we mention methods that propose to reformulate the elliptic problem into an equivalent one before its resolution. In this class of method, Del Castillo-Negrete and Chacón [dCNC11] introduced a method which avoids the discrete matrix inversion (usually ill-conditioned for highly anisotropic diffusion tensors) by using the Lagrangian Green's function. The method is based on an integral formulation for the parallel transport equation that eliminates spurious perpendicular diffusion. However, results are limited to constant diffusion values in the diffusion direction with a null perpendicular component (one-dimensional diffusion tensor). The *Asymptotic Preserving* (AP) methods belong to this category. They are designed to preserve, at the discrete level, the asymptotic limit between micro-macroscopic problems [Jin12]. Originally developed to capture the steady-state solution for neutron transport in diffusive regime in the later '80s, see Larsen *et al.* [LMJ87], applications to strong anisotropies when solving elliptic equation appeared in Degond *et al.* [DDN10], where the anisotropic direction is aligned with one coordinate. A generalized version was presented in [DDL$^+$12] for any given anisotropy vector director not aligned with the mesh and/or coordinates. Then, in [DLNN12] a reformulated version called Micro-Macro decomposition was presented. In Mentrelli *et al.* [MN12], the method modelled a simplified non-linear temperature equation model for magnetically confined plasmas, showing independence of the method on anisotropic strength. The application of the described methods is however limited to simple magnetic field, and seems to be not extensible to generalized magnetic fields. More recently, Narski *et al.* [NO14] introduced a stabilization term to conserve accuracy in (*magnetic islands*), generalizing the cited AP method to real magnetic field cases.

In the following, we will focus on solution methods proposed in the frame of finite-differences discretizations which are at the heart of this PhD work.

2.2 The mathematical model

Our focus lies in the resolution of strongly anisotropic diffusion problems using first-order, implicit time discretization, or anisotropic Poisson's equations occurring when investigating stationary solutions of the same diffusion problems. These two problems can be described by the following, generic, elliptic boundary value problem:

$$\begin{cases} -\boldsymbol{\nabla}\cdot\boldsymbol{\mathcal{K}}\boldsymbol{\nabla}T+\mu T=S & \text{in } \Omega,\\ \beta\nabla_{b\|}T+\gamma T=g & \text{on } \Gamma, \end{cases} \tag{2.1}$$

where Ω is a bounded domain in $\mathbb{R}^3$ with boundary Γ, provided with an orthonormal basis $(\mathbf{e}_x,\mathbf{e}_y,\mathbf{e}_z)$ associated to Cartesian coordinates (x,y,z). We assume that the variables of the problem satisfy the usual ellipticity and regularity assumptions. μ is a positive (or zero) constant, and S is a given source term. The coefficients β and γ, and g are used to define general boundary conditions, that can be of Dirichlet ($\beta=0$), Neumann ($\gamma=0$) or Robin ($\beta\neq 0,\gamma\neq 0$) type. With such a problem, both periodic and bounded domains can be considered that allows us to disconnect the discretization of the equation in the interior of the domain and at the boundaries.

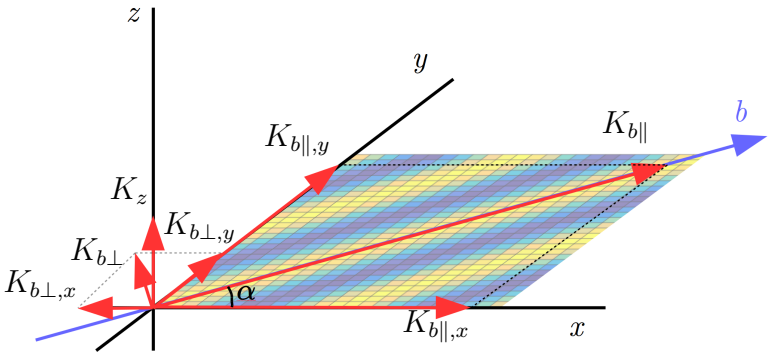


Figure 2.1: Chart of projected diffusion values (parallel and perpendicular direction of $\mathbf{b}$) on the Cartesian directions, with $K_{b\|}\gg K_{b\perp}$ and $K_{b\perp}\approx K_z$, Eq. 2.21.

The anisotropy of the problem in the 2D plane is taken into account via the definition of the symmetric diffusion tensor $\boldsymbol{\mathcal{K}}$, where the first eigenvalue $K_{b\|}$ (direction of anisotropy, called here $K_{b\|}$ in reference of fusion plasma) is assumed to fix the dominant diffusion direction ($K_{b\|}\gg K_{b\perp}$), that we can identify with the normalized eigenvector $\mathbf{b}$. $K_{b\perp}$ is the diffusion in the perpendicular direction. The latter is assumed to be isotropic with $K_{b\perp}\approx K_z$. This is illustrated on Fig. 2.1. Under these assumptions, the system reads:

$$-\boldsymbol{\nabla}\cdot\left[\boldsymbol{\mathcal{R}}\begin{bmatrix} K_{b_\|} & 0 & 0\\ 0 & K_{b_\perp^1} & 0\\ 0 & 0 & K_{b_\perp^2}\end{bmatrix}\boldsymbol{\mathcal{R}}^{-1}\right]\begin{Bmatrix}\partial T/\partial x\\ \partial T/\partial y\\ \partial T/\partial z\end{Bmatrix}+\mu T=S. \tag{2.2}$$

where $\boldsymbol{\mathcal{R}}$ defines a 3D rotation matrix.

Let's notice that $\boldsymbol{b}$ can be function of space. Gradients along the parallel and perpendicular directions are then defined as $\nabla_{b\|}=\boldsymbol{b}\cdot\boldsymbol{\nabla}$ and $\boldsymbol{\nabla}_{b\perp}=\boldsymbol{\nabla}-\boldsymbol{b}\nabla_{b\|}$, respectively.

Great simplifications can be obtained by defining an orthonormal basis constituted by the normalized eigenvectors of $\boldsymbol{\mathcal{K}}$, namely $(\mathbf{b},\mathbf{e}_\perp^1,\mathbf{e}_\perp^2)$, and the associated aligned

coordinate system $(b_{\|}, b^1_{\perp}, b^2_{\perp})$. The orthogonality of the eigenvectors of $\boldsymbol{\mathcal{K}}$ comes from its symmetry property. The problem Eq. (2.2) then reads:

$$-\boldsymbol{\nabla} \cdot \begin{bmatrix} K_{b_{\|}} & 0 & 0 \\ 0 & K_{b^1_{\perp}} & 0 \\ 0 & 0 & K_{b^2_{\perp}} \end{bmatrix} \begin{Bmatrix} \partial T/\partial b_{\|} \\ \partial T/\partial b^1_{\perp} \\ \partial T/\partial b^2_{\perp} \end{Bmatrix} + \mu T = S. \tag{2.3}$$

2.3 The finite-difference methods

Basically, finite-difference methods define derivatives at the discrete level from formulation obtained from Taylor expansion. Let f be an $(n+1)$ times differentiable function on an open interval containing the points x and $x+h$. Then:

$$f(x+h) = f(x) + hf'(x) + h^2 \frac{f''(x)}{2!} + ... + h^n \frac{f^{(n)}(x)}{n!} + R_{(n)}(x), \tag{2.4}$$

where

$$R_{(n)}(x) = h^{n+1} \frac{f^{(n+1)}(\zeta)}{(n+1)!}. \tag{2.5}$$

Being $f(x)$ and $f(x+h)$ two known function values in x and $x+h$ respectively, and $\zeta \in]x, x+h[$, the first-order derivative of f can be approximated in this interval using the Taylor expansion of f considering the three first terms of the expansion:

$$f'(x) = \frac{f(x+h) - f(x)}{h} + \frac{R(x)}{h}. \tag{2.6}$$

This expression is the *forward* finite-differences, where R is the residual term of the Taylor expansion. As seen in Eq. 2.5, the residual value in Eq. 2.6:

$$R_{(n)}(x) = -\frac{h}{2!} f''(\zeta) = \mathcal{O}(h), \tag{2.7}$$

being $\mathcal{O}(h)$ the truncation error. Eq. 2.4 can be rewritten for the interval $x-h$ to obtain the *backward* finite-differences. By differentiation of *forward* and *backward* expressions, we obtain the *centered* version:

$$f'(x) = \frac{f(x+h) - f(x-h)}{2h} + \mathcal{O}(h^2). \tag{2.8}$$

Note the *centered* version presents a residual term one order higher than previous versions due to the differentiation between the *forward* and *backward* versions eliminate the residual term of first order. More accurate $f'(x)$ approximations can be obtained knowing the function value in additional points.

For a second order derivative approximation of f, knowing $f(x-h)$, f_x and $f(x+h)$ in the interval $]x-h, x+h[$, and considering the approximation of the first derivative of Eq. 2.8, then, the centered second order derivative reads:

$$f''(x) = \frac{f(x-h) - 2f(x) + f(x+h)}{h^2} + \mathcal{O}(h^2). \tag{2.9}$$

Here, the residual term inherits the same precision than first order centered derivative, Eq. 2.8.

2.3.1 Grid definition and notation

The computational domain is the cube $[0, 2\pi] \times [0, 2\pi] \times [0, 2\pi]$ in the (x, y, z) directions, respectively. The grid is structured and uniform. Each cell in the grid can be addressed by indices (i, j, k), and each vertex has coordinates $x_i = i(2\pi/N_x)$, $y_j = j(2\pi/N_y)$, $z_k = k(2\pi/N_z)$ for $(i, j, k) \in [1, N_x] \times [1, N_y] \times [1, N_z]$, where N_x, N_y, N_z are the numbers of points in each direction. Distances between grid points are defined as $\Delta x = x_{i+1} - x_i$, $\Delta y = y_{j+1} - y_j$ and $\Delta z = z_{k+1} - z_k$. For clarity, (i, j, k) is also labelled by $\lambda = (i-1)\, N_y\, N_z + (j-1)\, N_z + k$ ($\lambda = (i-1)\, N_y + j$ in $x - y$ cases).

In the following, the discretization will be oriented, with $\boldsymbol{b}$ defining the local positive direction at any (i, j, k) point. Quantities to discretize may thus eventually be superscripted with $+$ or $-$ when needed. In the following, the set of values at the grid points, and at the points where fluxes are estimated will be denoted grid space (GS) and flux space (FS), respectively. All quantities belonging to FS will be superscripted by tilde$\tilde{}$ in Chapters 2 and 3.

2.3.2 Non aligned methods

In the following, methods based on stencils independent of the diffusion tensor and using differentiation formula in the x, y, and z directions (Eq. 2.2) will be denoted *non-aligned methods*. On the contrary, methods using stencils adapted to the direction of $\boldsymbol{b}$ (Eq. 2.3) will be denoted *aligned methods*.

The classic scheme

The classic conservative approach to the diffusion operator (asymmetric scheme in [GYKL05, vEKdB14]) is based in the lowest order finite difference calculus centered on the control volume faces. Given any grid point, a Control Volume (CV) is defined at the midpoint of neighbouring grid points (red dashed line in Fig. 2.2).

The gradient in the Cartesian basis is evaluated centered to CV faces. The function is interpolated with a polynomial approach as follows:

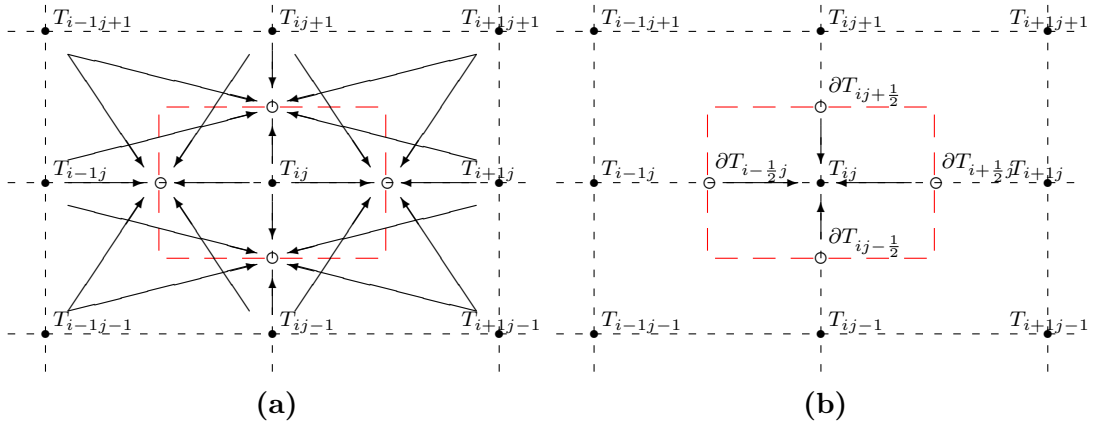


Figure 2.2: The classic scheme: (a) stencil points for the polynomial interpolation centered at CV faces. (b) the complete operator obtained from the gradient evaluation at the CV faces centered to the evaluated grid point.

$$
\begin{aligned}
T^{int}_{i+1/2,j} &= p_{i\,j}(x_{i+1/2}, y_j), \\
T^{int}_{i,j+1/2} &= p_{i\,j}(x_i, y_{j+1/2}),
\end{aligned}
\tag{2.10}
$$

where

$$
\begin{aligned}
p_{i+1/2,j}(x,y) =& a_{10} + a_{11}x + a_{12}y + a_{13}xy + a_{14}y^2 + a_{15}xy^2, \\
p_{i,j+1/2}(x,y) =& a_{20} + a_{21}x + a_{22}y + a_{23}xy + a_{24}x^2 + a_{25}yx^2.
\end{aligned}
\tag{2.11}
$$

It is analogous for the $(i-1/2\,,j)$ and $(i\,,j-1/2)$ faces. The values of the discrete gradients are obtained from these 2D Eqs. 2.10 by differentiation in each direction. In $i+1/2, j$ (analogous for each CV face), the derivatives lead to:

$$\partial_x p_{i+1/2,j} = a_{11} + a_{13}y + a_{15}y^2 \tag{2.12}$$

$$\partial_y p_{i+1/2,j} = a_{12} + a_{13}x + 2a_{14}y + 2a_{15}xy \tag{2.13}$$

Solving the linear system, the discrete derivatives lead to the finite-differences formula (analogous to the lowest order finite-differences obtained from Taylor expansion):

$$
\begin{aligned}
(\partial_x T)_{i+1/2,j} &\approx \frac{T_{i+1,j} - T_{i,j}}{\Delta x}, \\
(\partial_y T)_{i+1/2,j} &\approx \frac{T_{i+1,j+1} + T_{i,j+1} - T_{i+1,j-1} - T_{i,j-1}}{4\Delta y}, \\
(\partial_x T)_{i,j+1/2} &\approx \frac{T_{i-1,j} + T_{i-1,j+1} - T_{i+1,j+1} - T_{i+1,j}}{4\Delta x}, \\
(\partial_y T)_{i,j+1/2} &\approx \frac{T_{i,j+1} - T_{i,j}}{\Delta y}.
\end{aligned}
\tag{2.14}
$$

Considering uniform parallel diffusion, in the direction along vector $\boldsymbol{b}$, the diffusion components can be calculated in the control volume faces by an arithmetic mean, Eq. 2.15, or harmonic mean, Eq. 2.16. Both means of two values are very close for values of the same magnitude, but the harmonic mean is always bounded in absolute value by twice the absolute value of the smallest of the two numbers [ADLT06]. This fact makes the harmonic mean more robust facing non-linearity lead by turbulent flows:

$$K_{b\|,i+\frac{1}{2}j} = \frac{K_{b\| \, i+1j} + K_{b\| \, ij}}{2}, \tag{2.15}$$

$$\frac{2}{K_{b\| \, i+\frac{1}{2}j}} = \frac{1}{K_{b\| \, i+1j}} + \frac{1}{K_{b\| \, ij}}. \tag{2.16}$$

Assuming $K_{b\|}$ is uniform in the whole domain, for any point T_{ij}, the Classic discrete parallel diffusion operator leads to:

$$\begin{aligned} \nabla \cdot (K_{b\|} \cdot \nabla_{b\|} T_{ij}) \approx & \frac{K_{b\|} \cos^2 \alpha}{\Delta x^2}(T_{i-1j} - 2T_{ij} + T_{i+1j}) \\ & + \frac{K_{b\|} \sin^2 \alpha}{\Delta y^2}(T_{ij-1} - 2T_{ij} + T_{ij+1}) \\ + 2\frac{K_{b\|} \sin\alpha \, \cos\alpha}{\Delta x \Delta y}(-T_{i+1j-1} + T_{i+1j+1} - T_{i-1j+1} + T_{i-1j-1}). & \end{aligned} \tag{2.17}$$

The Günter's scheme

Günter *et al.* [GYKL05] proposed a 9-point stencil non-aligned approach centered in the grid nodes, Fig. 2.3(a), and using the same stencil than the classic scheme. The discrete derivative is evaluated as follows:

$$\begin{aligned} \partial_x T_{i+\frac{1}{2}j+\frac{1}{2}} &\approx \frac{T_{i+1j+1} + T_{i+1j} - T_{ij+1} - T_{ij}}{2\Delta x}, \\ \partial_y T_{i+\frac{1}{2}j+\frac{1}{2}} &\approx \frac{T_{ij+1} + T_{i+1j+1} - T_{ij} - T_{i+1j}}{2\Delta y}, \end{aligned} \tag{2.18}$$

It is analogous for the $(i-1/2, j+1/2)$, $(i+1/2, j-1/2)$ and $(i-1/2, j-1/2)$ locations. The parallel diffusion tensor component can be evaluated at the grid nodes reformulating the arithmetic mean, Eq. 2.15 and the harmonic mean, Eq. 2.16:

$$K_{b\| \, i+\frac{1}{2}j+\frac{1}{2}} = \frac{K_{b\| \, i+1j+1} + K_{b\| \, ij+1} + K_{b\| \, i+1j} + K_{b\| \, ij}}{4}, \tag{2.19}$$

$$\frac{4}{K_{b\| \, i+\frac{1}{2}j+\frac{1}{2}}} = \frac{1}{K_{b\| \, i+1j+1}} + \frac{1}{K_{b\| \, ij+1}} + \frac{1}{K_{b\| \, i+1j}} + \frac{1}{K_{b\| \, ij}}. \tag{2.20}$$

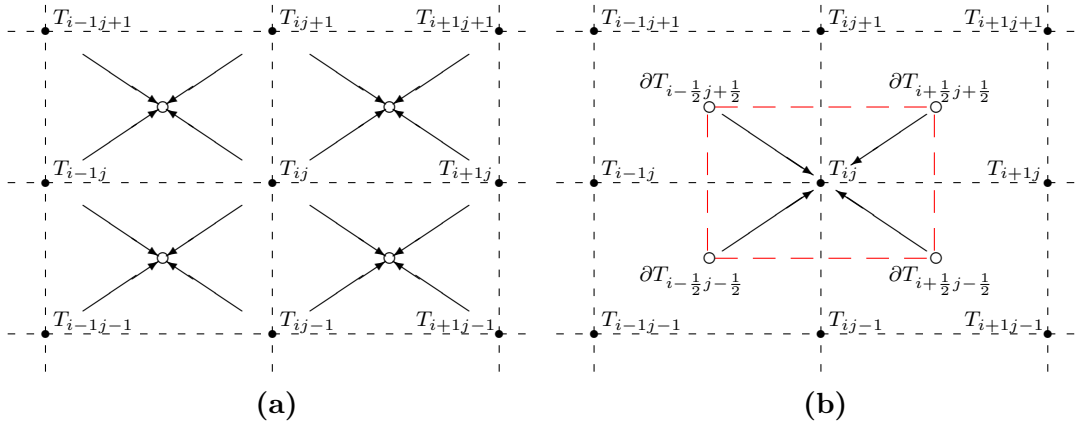


Figure 2.3: The Günter's scheme: (a) The gradient evaluations are obtained at the grid nodes from neighboring grid points, Eq. 2.18. In (b), the discrete centered diffusion operator is obtained from gradient evaluations at the CV corners.

From the tensor values, fluxes $q_{i+\frac{1}{2}j+\frac{1}{2}}$ are obtained taking the corner derivatives as follows:

$$q_{i+\frac{1}{2}j+\frac{1}{2}} = -\left[\boldsymbol{\mathcal{R}}\begin{bmatrix} K_{b\|\,i+\frac{1}{2}j+\frac{1}{2}} & 0 \\ 0 & 0 \end{bmatrix}\boldsymbol{\mathcal{R}}^{-1}\right]\begin{Bmatrix} \partial_x T_{i+\frac{1}{2}j+\frac{1}{2}} \\ \partial_y T_{i+\frac{1}{2}j+\frac{1}{2}} \end{Bmatrix}. \tag{2.21}$$

Finally, the divergence of the fluxes defines the full operator as follows:

$$\begin{aligned}\nabla\cdot q = &\frac{q_{x\,i+\frac{1}{2}\,j+\frac{1}{2}} + q_{x\,i+\frac{1}{2}\,j-\frac{1}{2}} - q_{x\,i-\frac{1}{2}\,j+\frac{1}{2}} - q_{x\,i-\frac{1}{2}\,j-\frac{1}{2}}}{2\Delta x} \\ &+\frac{q_{y\,i+\frac{1}{2}\,j+\frac{1}{2}} + q_{y\,i-\frac{1}{2}\,j+\frac{1}{2}} - q_{y\,i+\frac{1}{2}\,j-\frac{1}{2}} - q_{y\,i-\frac{1}{2}\,j-\frac{1}{2}}}{2\Delta y}.\end{aligned} \tag{2.22}$$

The Günter's symmetric discretization maintains some symmetry characteristics of the differential operator discretized: it preserves the self-adjointness of the latter (since it can be written in the Support Operator Method framework, SOM) at the discrete level, giving a positive semi-definite sparse matrix for the divergence of the flux. Aspects of the SOM are fully explained in Sec. 2.3.3.

The previous schemes can be implemented in any grid coordinates system (Cartesian, Cylindrical or Polar systems). In a general framework, the anisotropic tensor $\boldsymbol{\mathcal{K}}$ is not parallel to any of the previous grid coordinates: finite-differences of a characteristic anisotropic field (that is, strong field variations on the perpendicular direction combined with slow variations on the parallel direction). This fact requires a high number of grid points to reach a satisfactory numerical precision.

The use of an aligned reference simplifies the treatment of this differential operator in the way no rotation is required.

2.3.3 Aligned schemes

Introduction

The introduction on aligned methods is focused on plasma modelling and simulation.

Roberts *et al.* [RT65] introduce the *twisted* coordinate system, based on the track of unperturbed magnetic field lines in the study of resistive instabilities of a fluid supported by a sheared magnetic field. In Dewar *et al.* [DG83], a coordinate set is established to transform the toroidal magnetic field lines into *straight* magnetic field lines, under the assumption of nested toroidal magnetic surfaces in aligned grids. The same assumptions are given in Cowley *et al.* [CKS91] and later by Hammet *et al.* [HBD+93]. In general, all the previous approaches are based on 2D simplification, considering magnetic surfaces aligned to the grid points for a given radial position, Fig. 2.4a.

Scott [Sco01] relieved the constraint of aligning the mesh by introducing a local coordinate system on each grid point, b_{ij} in Fig. 2.4b. This approach avoids grid deformations by using the so-called "shifted metric" procedure, taking into account the Hamada [Ham62] flux tube approach (global aligned coordinates), and splitting it into local shifted tubes related to grid points.

In Ottaviani [Ott11], a set of local aligned coordinates is presented considering the flute property of plasma turbulent flows: it opens the possibility to a grid points reduction in a chosen direction, providing enough information on the fine structure by the variation of any other direction, Fig. 2.4c. A full review of aligned coordinates considered in Hammet *et al.* [HBD+93] and Scott [Sco01] is compared with Ottaviani approach [Ott11] (renamed Flux-Coordinate Independent, FCI, for the usual tokamak cylindrical coordinates) in Hariri *et al.* [HO13].

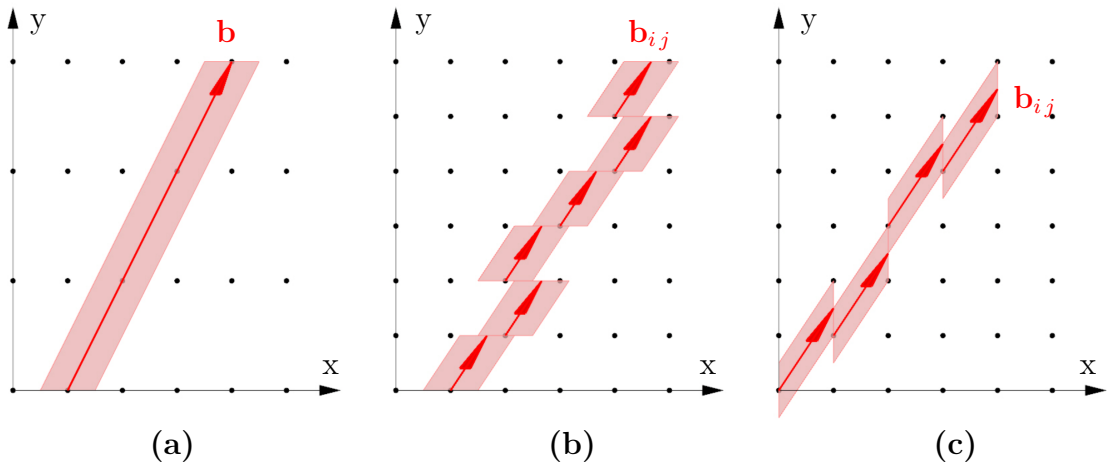


Figure 2.4: Representation of historic evolution of aligned coordinates by (a) Arakawa [Ara97] scheme, (b) Scott [Sco01] shifted scheme and (c) Ottaviani [Ott11] shifted scheme.

Geometrical quantities definition

Aligned methods are defined using geometrical quantities obtained from the parallel diffusion line $\boldsymbol{b}$ trace. This section defines the main quantities used later in finite-differences equations, being $\boldsymbol{b}$ known throughout the domain Ω. For simplicity, we assume $\boldsymbol{b}$ parallel and uniform in Ω.

First quantity are the geometrical intersection between $\boldsymbol{b}$ defined in the grid point $ijk \in \mathcal{X}_i$ plane, with the neighboring plane $\mathcal{X}_{i+1}$ at (x_{i+1}, y^+, z^+) (Fig. 2.5):

$$y^+ = y + \int_{x_i}^{x_{i+1}} \frac{b_y}{b_x} dx, \quad z^+ = z + \int_{x_i}^{x_{i+1}} \frac{b_z}{b_x} dx, \tag{2.23}$$

being b_x, b_y and b_z the components of $\boldsymbol{b}$ in the Cartesian frame (x, y, z). The expression for y^- and z^- directions are obtained replacing the integral limits in Eq. 2.23. The relative lengths are defined as $\delta y^+ = y^+ - y$ and $\delta z^+ = z^+ - z$ in Fig. 2.5.

Another useful quantity is the length of $\boldsymbol{b}$ curve between two adjacent planes $\mathcal{X}_i$ and $\mathcal{X}_{i+1}$, defined as $d_{b\|}$:

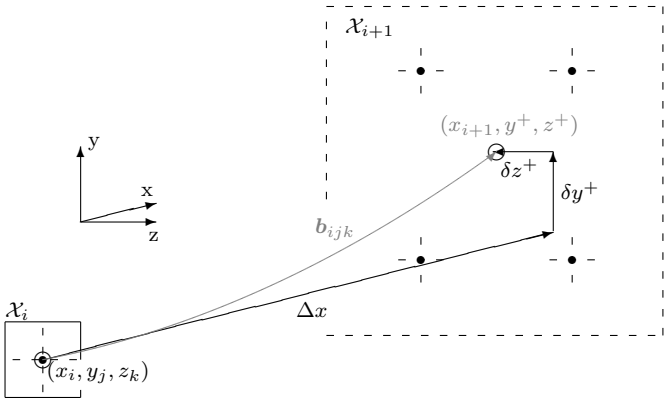


Figure 2.5: Parallel diffusion vector tracking chart from $(x, y, z) \in \mathcal{X}_{i+}$ to $(x, y^+, z^+) \in \mathcal{X}_{i+1}$. δy^+ and δz^+ correspond to the integral part of Eq. 2.23 for + index, and its corresponding part in the equation for z^+, respectively.

$$d_{b\|} = \left| \int_{x_i}^{x_{i+1}} \frac{\sqrt{b_x^2 + b_y^2 + b_z^2}}{|b_x|} \, \mathrm{d}x \right|, \tag{2.24}$$

For simplicity, $\boldsymbol{b}$ can be considered straight, uniform and defined in $x - y$ plane. Under this assumption, Eqs. 2.23, 2.24, becomes:

$$d_{b\|} = \frac{\Delta x}{\cos \alpha}; \qquad y^+ = y + \Delta x \tan \alpha, \tag{2.25}$$

where α is the pitch angle, defined by $\boldsymbol{b}$ slope in Ω:

$$\tan\alpha = \frac{b_y}{b_x} = \frac{\delta y^+}{\Delta x}. \tag{2.26}$$

A basic aligned discretization

To introduce the aligned approach, the same CV of Classic and Günter is considered to discretize Eq. 2.3 (Fig. 2.6). The interpolation step is done here with the same expressions of the classic approach, Eqs. 2.10, now interpolating on the CV intersection with the parallel diffusion line $K_{b\parallel}$ for the parallel dynamics, Eqs. 2.27. The following process is analogous for the perpendicular direction $K_{b\perp}$.

$$\begin{aligned}
(x^+_{int}, y^+_{int}) &= \left(\frac{\Delta x}{2}, \frac{\Delta x}{2} atan(\alpha)\right), \quad when\ y^+_{int} < \frac{\Delta y}{2} \\
(x^+_{int}, y^+_{int}) &= \left(\frac{\Delta y}{2} atan(\pi/2 - \alpha), \frac{\Delta y}{2}\right), \quad when\ y^+_{int} > \frac{\Delta y}{2} \\
(x^+_{int}, y^+_{int}) &= \left(\frac{\Delta x}{2}, \frac{\Delta y}{2}\right), \quad when\ y^+_{int} = \frac{\Delta y}{2}
\end{aligned} \tag{2.27}$$

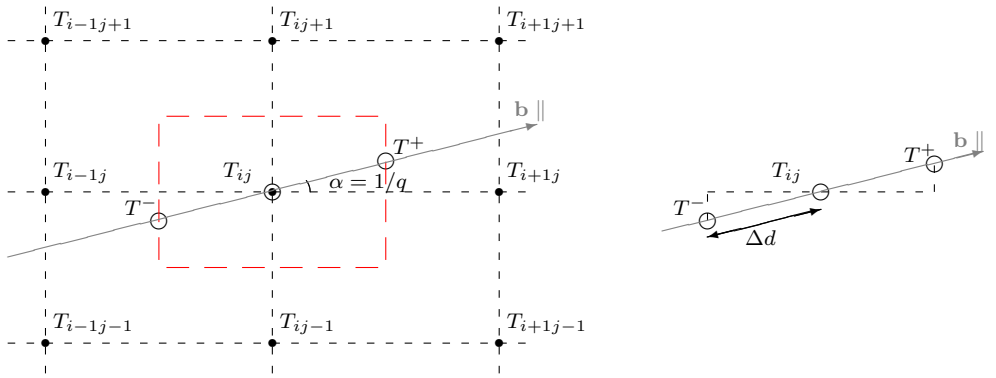


Figure 2.6: (a) Interpolation coordinates obtained from the intersection between the CV and the parallel diffusion line. (b), final finite-differences stencil parallel to $\boldsymbol{b}$.

The finite-differences steps can be done in one stage now. The discrete diffusion operator reads:

$$\nabla \cdot (K_{b\parallel} \cdot \nabla_{b\parallel} T_{ij}) \approx \frac{K_{b\parallel\, ij}}{\Delta d^2} \left(T^+ - 2T_{ij} + T^-\right), \tag{2.28}$$

where T^+ and T^- represents the interpolated field values, and Δd the distance of the parallel diffusion direction between the T_{ij} coordinates and the interpolated values:

$$\Delta d = \frac{x(T^+) - x(T_{i,j})}{\cos\alpha} = \frac{y(T^+) - y(T_{i,j})}{\sin\alpha}. \tag{2.29}$$

Ottaviani's scheme

Since this approach is a suitable approximation to solve Eq. 2.3, the location of the interpolated points on the parallel direction $\boldsymbol{b}$ limit the possibility to a better approximation in terms of precision and optimized grids for highly anisotropic flows. Ottaviani *et al.* [Ott11, HO13] propose the already cited FCI scheme as aligned scheme. The method considers an interpolation step in two adjacent planes located in $x_{i\pm1}$ to the grid point T_{ij}, Fig. 2.7.

This approach reduces the interpolation step to one direction in 2D cases (y-direction in Fig. 2.7) or a plane in 3D cases. The main advantage of this method is the possibility to adapt the grid resolution to fields with flute property, where highly anisotropic diffusion damp non-parallel modes rapidly, leading to elongated structures in the parallel direction with strong field variations in the perpendicular direction (rapid slope variations).

Ottaviani scheme is an adaptation of the discrete operator to the field: one direction (2 directions in 3D cases) of the grid provide all the information of fine structure, being the second (third) one drastically reduced to represent coarse structures. For example, in toroidally confined plasma, turbulent structures are elongated in the toroidal direction, having strong field variations in density and electric potential in radial-poloidal directions. This orientation is related to the magnetic field line pitch angle with respect to the toroidal direction, which is generally small ($\alpha \lesssim 15$ degrees) [PKFC$^+$17] .

In 2D Cartesian coordinates, the orientation of structures solving Eq. 2.3 depends on the parallel direction. By analogy with the previously described for plasma, the pitch angle, Eq. 2.26, is considered small ($\alpha > \pi/4$). Therefore, the field is characterized by a strong variation in y direction, and a slow variation in x.

Then, Ottaviani proposes to reduce the interpolation step to the intersection of $\boldsymbol{b}$ in the surrounding planes. Here, a 2D version of Ottaviani approach is described, considering the interpolation step on y-direction, and the finite-differences step a x-function. In Fig. 2.7, T^{int+} and T^{int-} are obtained by linear interpolation:

$$T^{int+} = (1-f)\,T_{i+1,j} + f\,T_{i+1,j+1}, \tag{2.30}$$

where f represent a linear interpolation factor in function of the grid spacing and the intersection of $K_{b\|ij}$ in $i \pm 1$. Considering the assumptions for Eqs. 2.25 2.26, Eq. 2.30 leads:

$$T^{int+} = \left(1 - \frac{\Delta x \tan\alpha}{\Delta y}\right) T_{i+1,j} + \frac{\Delta x \tan\alpha}{\Delta y}\, T_{i+1,j+1}. \tag{2.31}$$

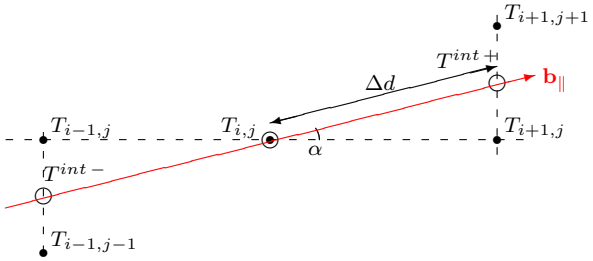


Figure 2.7: Ottaviani discretization interception the surrounding planes in $x_{i\pm1}$. T^{int+} and T^{int-} are interpolated values in y-direction.

The discrete parallel gradient is defined by finite-differences in the mid-plane as follows (analogous for $\nabla_{b\parallel}T_{i-1/2j}$):

$$\nabla_{b\parallel}T_{i+1/2j} \approx \frac{T^{int+} - T_{ij}}{\Delta d^{+}}. \tag{2.32}$$

Then, the complete parallel diffusion operator, is solved centered in T_{ij} by finite-differences as follows (considering here $\Delta d^{+} = \Delta d^{-} = \Delta d$):

$$\nabla \cdot (K_{b\parallel} \cdot \nabla_{b\parallel}T_{ij}) \approx \frac{K_{b\parallel\, ij}}{\Delta d^2}\left(T^{int+} - 2T_{ij} + T^{int-}\right). \tag{2.33}$$

Fine structures here can be defined in y direction, being the parallel modes represented by N_x resolution: this formulation opens to a large of N_y/N_x relation adapted to anisotropic structures.

The support operator method (SOM)

The expression for the parallel gradient above now enables the use of the support-operator method (SOM) [MS08, SS94, MRS98, LMS14, MSS00] to obtain the parallel Laplacian. For any function T in H^2, and Ψ in H^1, both countinuous in Ω, and with suitable boundary conditions (bi-periodic or homogeneous Dirichlet), we establish the temporal conservation law of the quantity T as variation across the domain boundary as:

$$\frac{dT}{dt} = -\oint_S (\mathbf{n}, \mathbf{m}) dS, \tag{2.34}$$

where the quantity T can be seen as the divergence of the quantity u in the domain volume:

$$T = \int_V u\, dV, \tag{2.35}$$

the total amount of quantity u, being the flux $\mathbf{n}$ across the boundary:

$$\mathbf{n} = -\mathcal{K}\,\nabla u. \tag{2.36}$$

The integral identity of the conservation law, Eq. 2.34, leads:

$$\frac{dT}{dt} = \frac{d}{dt}\int_V u dV = -\int_V \boldsymbol{\nabla}\cdot\mathbf{n} dV = -\oint_S (\mathbf{n},\mathbf{m}) dS, \tag{2.37}$$

where the equality obtained establish:

$$-\int_V \boldsymbol{\nabla}\cdot\mathbf{n} dV = -\oint_S (\mathbf{n},\mathbf{m}) dS. \tag{2.38}$$

In general, $-\boldsymbol{\nabla}\cdot(\mathcal{K}\cdot\nabla)$ operator is positive definite and self-adjoint. This property is derived from the following integral identity:

$$\int_V \Psi\boldsymbol{\nabla}\cdot\mathbf{n} dV + \int_V (\mathbf{n},\nabla\Psi) dV = \oint_S \Psi(\mathbf{n},\mathbf{m}) dS. \tag{2.39}$$

Eq. 2.39 establish the connection between gradient and divergence operators as the anti-adjoint one to each other.

This relationship can be reproduced at the discrete level to obtain a self-adjoint and positive definite operator on this level. According to [SS94], SOM method on the discrete level process can be summarized in 5 points:

- Differential equation in terms of the invariant first-order differential operator gradient and divergence;
- Select in the grid level the scalar and vector functions to be located.
- Define one of the first order operators, gradient or divergence, as a *prime* operator.
- Discretization of the prime operator.
- The remaining operator, called *derived* operator, is obtained from the discretization of the prime operator and a difference analog of the integral identity, Eq. 2.39.

The discrete analog described here is applied by Stegmeir er al. [SCM+16, SMC+17] (described in the following section) and also in the proposed method described in Chapter 3.

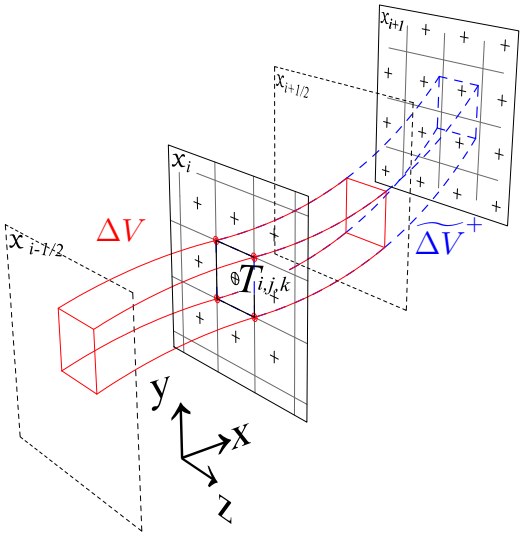


Figure 2.8: Stegmeir CV definition for $\nabla_{b\|}T$(blue slashed lines) and the local grid point (red line).

The Stegmeir's scheme

In Stegmeir *et al.* [SCM+16, SMC+17] the Field Line Map (FLM) approach is presented. Based on the magnetic field lines trace geometry, Stegmeir and co-authors present an integration method for the gradient operator from interpolated field values in the magnetic field lines, before deriving the full diffusion operator with SOM. The resulting approach presents lower numerical diffusion than the physical perpendicular diffusion given by the Arakawa discretization, higher convergence tendency and a reduction of the number of unknowns as in [Ott11].

The method propose a parallel discretization obtaining the gradient calculus by integration method:

$$\nabla_{b\|}T = \frac{\mathbf{b}}{\|\mathbf{b}\|}\cdot\nabla T = \frac{1}{\|\mathbf{b}\|}\nabla\cdot(T\boldsymbol{b}) \equiv \lim_{V(K)\to 0}\frac{1}{\|\mathbf{b}\|V(K)}\int_S T\boldsymbol{b}\cdot\boldsymbol{n}dS, \tag{2.40}$$

being here V(K) the volume of the K^{th} Control Volume $\widetilde{CV}$ ($\widetilde{\Delta V^+}$ in Fig. 2.8). Eq. 2.40 is mimicked on the discrete level by flux boxes defined by surrounding magnetic field lines centered in the grid cells. This definition reduces the contributions to the toroidal ends of the flux box (here described for + position flux):

$$(\widetilde{\nabla_{b\|}T})_p^+ = \frac{1}{\widetilde{\Delta V_p^+}}\left(T^{int}_{i+1\,p}\,a_{i+1\,p}\,\boldsymbol{b}_{i+1\,p}\cdot\boldsymbol{n}_{i+1\,p} + T^{int}_{i\,p}\,a_{i\,p}\,\boldsymbol{b}_{i\,p}\cdot\boldsymbol{n}_{i\,p}\right), \tag{2.41}$$

where $(\widetilde{\nabla_{b\|}T})_p^+$ is geometrically obtained in the center of the $\widetilde{CV}$. Generalizing for all

points, a matrix $Q^{\pm}$ can be assembled to obtain p-gradients centered in $\widetilde{CV}_p$:

$$(\widetilde{\nabla_{b\|}T})^{+}_{p} = \sum_{\lambda} Q^{+}_{p\lambda} T_{\lambda}, \tag{2.42}$$

$Q^{\pm}$ is seen here as an application to obtain all p-fluxes from Grid points Space (GS) in the Flux Space (FS):

$$Q^{+}_{p\lambda} : GS \longrightarrow FS \tag{2.43}$$

According to SOM definition described in the previous section, gradient discretization is considered here the *prime* operator. Then, divergence can be obtained mimicking Eq. 2.39 assuming bi-periodic domain and/or the quantities vanish at the boundaries:

$$\int_{V} \Psi \boldsymbol{\nabla} \cdot \mathbf{n} dV = -\int_{\widetilde{V}} (\mathbf{n}, \nabla\Psi) \widetilde{dV}. \tag{2.44}$$

Divergence is the *derived* operator from the gradient discretization:

$$\langle \nabla \cdot \nabla_{b\|} T, \Psi \rangle_{GS} = -\langle \nabla_{b\|} T, \nabla_{b\|} \Psi \rangle_{FS}, \tag{2.45}$$

and at the discrete level:

$$\langle D^{+}_{b\|} T, \Psi \rangle_{GS} = -\langle [K_{b\|}] Q^{+} T, Q^{+} \Psi \rangle_{FS^{+}}, \tag{2.46}$$

where $D^{+}_{b\|} T = \boldsymbol{\nabla} \cdot (K_{b\|} \cdot \nabla T)$ defined from Q^{+}, obtaining an parallel diffusion operator as an application in GS:

$$D^{+}_{b\|} : GS \longrightarrow GS. \tag{2.47}$$

Then, according with Eqs. 2.42 and 2.46, the integral on discrete level gives:

$$\langle D^{+}_{b\|} T, \Psi \rangle_{GS} \approx \sum_{\sigma} (\nabla \cdot [K] \nabla_{b\|} T)_{\sigma} \Psi_{\sigma} \Delta V_{\sigma}, \tag{2.48}$$

$$-\langle [K_{b\|}] Q^{+} T, Q^{+} \Psi \rangle_{FS^{+}} \approx -\sum_{p} \widetilde{\Delta V_p} \left(K_{b\| p} \sum_{\lambda} Q^{+}_{p\lambda} T_{\lambda} \right) \left(\sum_{\mu} Q^{+}_{p\mu} \Psi_{\mu} \right). \tag{2.49}$$

Here μ denotes the flux construction in FS. The equality of Eqs. 2.48 and 2.49 leads to the discrete diffusion operator:

$$D^{+}_{b\|} T = -\sum_{p} \frac{\widetilde{\Delta V_p}}{\Delta V_p} [Q^{+}_{p}]^{T} K_{b\| p} Q^{+}_{p} T, \tag{2.50}$$

Which shows the self-adjointness property of the operator in the discrete level. The complete operator is obtained considering also the "-" flux trough an analog discretization process. Then, the complete diffusion operator becomes:

$$D_{b\|}T = \frac{1}{2}(D^{+}_{b\|}T + D^{-}_{b\|}T). \tag{2.51}$$

Chapter 3

A new conservative finite-difference scheme for anisotropic elliptic problems in bounded domain

3.1 Introduction

This chapter introduces the new finite-difference scheme developed during the thesis. Based on interpolations aligned along the direction of the anisotropy, it is proved to be robust and conservative, and able to deal accurately with various boundary conditions as well. The chapter details first the scheme in terms of conservative discretizations of the parallel and perpendicular operators in the interior of the domain, Sec. 3.2. Then, a new compatible approach to deal with bounded problems is proposed in Sec. 3.4. Finally, numerical results based on analytical solutions show the accuracy and efficiency of the scheme for solving 2D elliptic problems, in both bi-periodic and bounded domain, including the case of the Poisson's equation with Robin boundary conditions, Sec. 3.5.

3.2 New conservative finite-difference scheme

The grid definition and notations are those introduced in Sec. 2.3.1.

3.2.1 Discretization of the parallel gradient $\nabla_{b\|}$

$\boldsymbol{b}$ can be function of space, and it is assumed here to be divergence-free. Let's notice that in problems where $\mathbf{b}$ would not be divergence-free, a divergence-free vector field everywhere co-linear to $\boldsymbol{b}$ could be constructed since the normalization of $\boldsymbol{b}$ is nowhere used in the problem. Under this assumption, and considering a control volume K of

volume V and surface S, the parallel gradient $\nabla_{b\|}$ can be estimated from the flux through S using the following definition for each control volume:

$$\nabla_{b\|}T = \frac{\mathbf{b}}{\|\mathbf{b}\|}\cdot\nabla T = \frac{1}{\|\mathbf{b}\|}\nabla\cdot(T\boldsymbol{b}) \equiv \lim_{V(K)\to 0}\frac{1}{\|\mathbf{b}\|V(K)}\int_S T\boldsymbol{b}\cdot\boldsymbol{n}dS \tag{3.1}$$

The control volume K around each grid point (i,j,k) is defined by the polygon with corners $(i, j\pm\frac{1}{2}, k\pm\frac{1}{2})$ in the $y-z$-plane $\mathcal{X}_i$, and extruded along the parallel direction up to the planes $\mathcal{X}_{i\pm\frac{1}{2}}$ (Fig. 3.1). At these planes, it partially overlaps neighbouring control volumes defined from grid points located in the adjacent planes $\mathcal{X}_{i+1}$ and $\mathcal{X}_{i-1}$. In the following, we will only consider by simplicity neighbouring control volumes defined in $\mathcal{X}_{i+1}$, the discretization being similar for control volumes defined in $\mathcal{X}_{i-1}$.

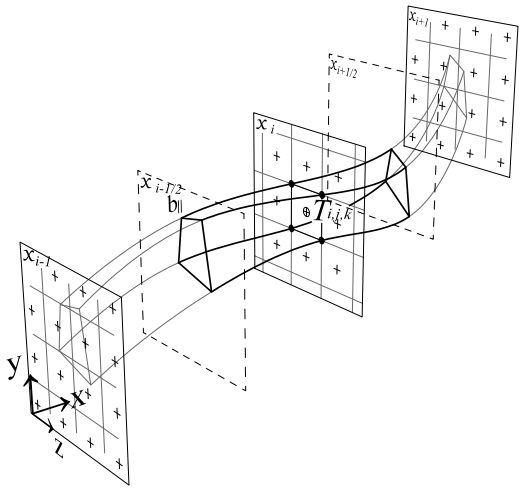


Figure 3.1: Sketch of a control volume (bold lines) defined in the grid space (GS) around $T_{i\,j\,k} \in \mathcal{X}_i$, and between $\mathcal{X}_{i\pm\frac{1}{2}}$ planes. General case with $\boldsymbol{b}$ as a function of (x,y,z).

The contact surfaces between control volumes and its neighbors are denoted a_p, $p = 1,...,N$, N being the total number of contact areas between two adjacent $\mathcal{X}$ planes (Fig. 3.2a). For each contact surface a_p, we consider the line that passes through its barycenter and follows the parallel direction, as illustrated in Fig. 3.2b. It intercepts the two planes $\mathcal{X}_i$ and $\mathcal{X}_{i+1}$ at two points of coordinates (x_i, y^-, z^-) and (x_{i+1}, y^+, z^+), where $(y^\pm, z^\pm)$ are defined between $(x_{i+1/2})$ and (x_{i+1}) in the direction of $\boldsymbol{b}$ $(+)$ (see a sketch on Fig. 3.3) or $(x_{i+1/2})$ and (x_i) in the opposite direction $(-)$ as:

$$y^+ = y + \int_{x_{i+1/2}}^{x_{i+1}}\frac{b_y}{b_x}dx, \quad y^- = y + \int_{x_{i+1/2}}^{x_i}\frac{b_y}{b_x}dx, \tag{3.2}$$

being b_x, b_y the components of $\boldsymbol{b}$ in the Cartesian frame (x,y,z). The expression for z^+ and z^- directions are obtained replacing b_y by b_z in Eq. 3.2.

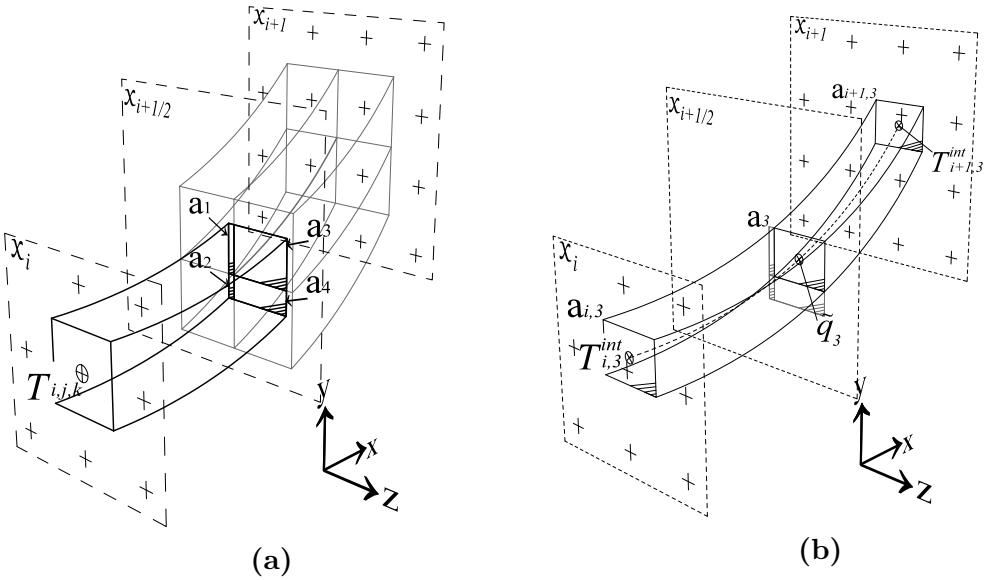


Figure 3.2: (a) Example of control volumes that overlap with contact surfaces a_p for $p = 1, ..., 4$. (b) Each overlapped surface allows to define a control volume in the flux space (FS), denoted $\widetilde{\text{CV}}$. Quantities used to evaluate the parallel flux Q_3 at $\mathcal{X}_{i+1/2}$ through the specific surface a_3 are included in the figure. Here, $\boldsymbol{b}$ is a function of (x) only for simplicity.

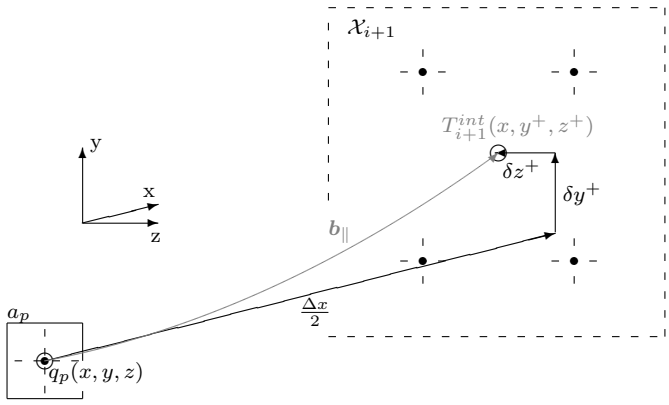


Figure 3.3: Parallel diffusion vector tracking chart from $(x, y, z) \in \mathcal{X}_{i+1/2}$ to $(x, y^+, z^+) \in \mathcal{X}_{i+1}$. δy^+ and δz^+ correspond to the integral part of Eq. 3.2 for + index, and its corresponding part in the equation for z^+, respectively.

The field values at these points are obtained by interpolation of the field values at the surrounding points with functions $f^{int}_{i\,p}$ and $f^{int}_{i+1\,p}$ in the corresponding planes. We can write for each p :

$$T^{int}_{i\,p} = f^{int}_{i\,p}\left(\{T_{i\,j\,k}\}_{i\,j=1,\dots N_y\;k=1,\dots N_z}\right). \tag{3.3}$$
$$T^{int}_{i+1\,p} = f^{int}_{i+1\,p}\left(\{T_{i+1\,j\,k}\}_{i\,j=1,\dots N_y\;k=1,\dots N_z}\right). \tag{3.4}$$

Thus, the parallel gradient $\widetilde{(\nabla_\parallel T)_p}$ can be discretized by approximating Eq. 3.1 as follows:

$$\widetilde{(\nabla_{b\parallel} T)_p} = \frac{1}{\widetilde{\Delta V_p}}\left(T^{int}_{i+1\,p}\, a_{i+1\,p}\, \boldsymbol{b}_{i+1\,p}\cdot\boldsymbol{n}_{i+1\,p} + T^{int}_{i\,p}\, a_{i\,p}\, \boldsymbol{b}_{i\,p}\cdot\boldsymbol{n}_{i\,p}\right), \tag{3.5}$$

where $\widetilde{\Delta V_p}$ is the volume obtained by integrating the surface a_p along the parallel direction between $\mathcal{X}_i$ and $\mathcal{X}_{i+1}$, and $\boldsymbol{n}_p$ the normal vector to the relevant surface. It is convenient to define a position at which the flux is evaluated, defined by the triplet $(\tilde{x}_p, \tilde{y}_p, \tilde{z}_p)$ which are the coordinates of the barycenter of the surface a_p.

The discretization of the parallel gradient (Eq. 3.5) defines a linear map Q from the space of grid values (GS) into the space of flux values (FS):

$$\begin{aligned} Q: \quad \mathrm{GS} &\to \mathrm{FS} \\ \{T\} &\to \{\widetilde{\nabla_{b\parallel} T}\} \end{aligned} \tag{3.6}$$

so that gradient values are given by:

$$\widetilde{(\nabla_{b\parallel} T)_p} = \sum_\lambda Q_{p\lambda} T_\lambda, \tag{3.7}$$

It is important to note here that the computation of left and right contributions of Eq. 3.5 are constructed to satisfy at the discrete level the fact the flux of $\boldsymbol{b}$ across the surface of any closed volume is zero, i.e.:

$$a_{i\,p}\boldsymbol{b}_{i\,p}\cdot\boldsymbol{n}_{i\,p} + a_{i+1\,p}\boldsymbol{b}_{i+1\,p}\cdot\boldsymbol{n}_{i+1\,p} = 0 \tag{3.8}$$

This ensures that Q is locally nilpotent for any constant temperature field, i.e. $\sum_\lambda Q_{p\lambda}T_\lambda = 0$ for all p if T_λ is constant. This property is crucial to the conservativity of the scheme. It is also noteworthy that this can be achieved in general by a consistant discretization only if the vector field $\boldsymbol{b}$ is divergence-free.

3.2.2 Discretization of the parallel Laplacian $\nabla \cdot K_{b_\parallel} \nabla_{b\parallel}$

The expression for the parallel gradient above now enables the use of the support-operator method (SOM) [MS08, SS94, MRS98, LMS14, MSS00] to obtain the parallel Laplacian, as found in Stegmeir et al. [SCM+16] in highly anisotropic diffusion. For any function $T \in \mathcal{H}^2$, and $\Psi \in \mathcal{H}^1$, both continuous in Ω, the Green formula reads:

$$\int_\Omega (\boldsymbol{\nabla} \cdot \mathbf{u}) \Psi dV + \int_\Omega \mathbf{u} \cdot \nabla \Psi dV = \int_\Gamma (\Psi \mathbf{u}) \cdot \mathbf{n} dS. \tag{3.9}$$

Considering $\mathbf{u} = -\mathcal{K} \nabla T$, the Green formula connects the gradient and the divergence operators. According to the definition of the $\mathcal{L}^2$-inner product in both scalar field and vector spaces H and $\mathbf{H}$, respectively, we write Eqs. 3.10, 3.11 for each one:

$$\langle -\nabla \cdot \mathcal{K} \nabla T, \Psi \rangle_H = -\int_\Omega \boldsymbol{\nabla} \cdot (\mathbf{u} \Psi) dV + \int_\Gamma (\Psi \mathbf{u}) \cdot \mathbf{n} dS. \tag{3.10}$$

$$\langle -\mathcal{K} \nabla T, \nabla \Psi \rangle_{\mathbf{H}} = \int_\Omega (\mathbf{u} \cdot \nabla \Psi) dV, \tag{3.11}$$

Eq. 3.9 establishes the connection between gradient and divergence operators as the self-adjoint one to each other. Considering any suitable boundary conditions (bi-periodic or homogeneous Dirichlet), the Green formula allows us to define the parallel diffusion operator directly from the parallel gradient as:

$$\langle -\nabla \cdot (\mathbf{b} \mathcal{K} \nabla_{b\|} T), \Psi \rangle = \langle \mathcal{K} \nabla_{b\|} T, \nabla_{b\|} \Psi \rangle, \tag{3.12}$$

Even if Eq. 3.12 is unambiguous at the continuous level, it involves two inner products, one defined in GS (Eq. E.18), and the other one in the FS (Eq. E.19) for any functions f and g as:

$$\langle f, g \rangle_{\mathrm{GS}} = \sum_\lambda f_\lambda \, g_\lambda \, \Delta V_\lambda, \tag{3.13}$$

$$\langle f, g \rangle_{\mathrm{FS}} = \sum_p \widetilde{f}_p \, \widetilde{g}_p \, \widetilde{\Delta V_p}. \tag{3.14}$$

According to Eq. 3.7, the inner product in FS can be estimated at the discrete level using evaluations of the diffusion on flux points denoted by $\{K_{b_\| \, p}\}$ as:

$$\langle [K] \nabla_{b\|} T, \nabla_{b\|} \Psi \rangle_{\mathrm{FS}} \approx \sum_p \widetilde{\Delta V_p} \left(K_{b\| \, p} \sum_\lambda Q_{p \, \lambda} T_\lambda \right) \left(\sum_\mu Q_{p \, \mu} \Psi_\mu \right). \tag{3.15}$$

Depending on the number of contact surfaces a_p, a certain number of flux values can be associated for each λ. In terms of the SOM formalism [SS94], $Q_{p\lambda}$ of Eq. 3.5, defined in FS is here the *prime operator*. The discretization of the divergence (*derived operator* in terms of SOM) defined in FS into GS is the adjoint of Q obtained by discrete analog of Eq. 3.12. Then, the full operator $(\nabla \cdot [K] \nabla_\|)$ is endomorphic in GS. The left-hand side of Eq. 3.12 leads at the discrete level to:

$$\langle \nabla \cdot [K] \nabla_{b\|} T, \Psi \rangle \approx \sum_\sigma (\nabla \cdot [K] \nabla_{b\|} T)_\sigma \Psi_\sigma \Delta V_\sigma. \tag{3.16}$$

According now to Eqs. 3.12, 3.15 and 3.16, one deduces by identification that:

$$-(\nabla \cdot [K]\nabla_{b\|}T)_\lambda \approx \frac{1}{\Delta V_\lambda}\sum_p \left(K_{b\| p} Q_{p\lambda} \sum_\mu Q_{p\,\mu} T_\mu \widetilde{\Delta V_p} \right), \tag{3.17}$$

The sum on μ term denotes the construction of the fluxes in FS from GS. Eq. 3.17 leads to:

$$(\nabla \cdot [K]\nabla_{b\|}T) \approx -\Delta V^{-1} Q^T \widetilde{[K]}\widetilde{\Delta V_p}\, Q\, T \tag{3.18}$$

Upon multiplication by the cell volume ΔV, the SOM provides a symmetric discrete matrix: the product of an operator and its adjoint is a self-adjoint positive definite operator, which maintains the symmetry of the matrix.

$$\mathcal{A}_{\lambda\mu}\Delta V_\lambda = \sum_p Q_{p\lambda} Q_{p\mu} [K]_p \widetilde{\Delta V_p} = \sum_p Q_{p\mu} Q_{p\lambda} [K]_p \widetilde{\Delta V_p} = \mathcal{A}_{\mu\lambda}\Delta V_\mu, \tag{3.19}$$

where $\widetilde{[K]}_p\widetilde{\Delta V_p} = \widetilde{[K\Delta V]}$ is a diagonal square matrix.

Finally, the conservativity of the scheme is verified by taking the special case $\Psi = 1$:

$$-\langle \nabla \cdot [K]\nabla_{b\|}T, 1\rangle_{\mathrm{GS}} = \langle [K]QT, Q1\rangle_{\mathrm{FS}} = 0 \tag{3.20}$$

which is equivalent to say that the average of the parallel Laplacian over the computational domain is zero. This property follows from Eq. 3.8, which implies $Q \cdot 1 = 0$, and leads to conservativity:

$$\langle (1 - \nabla \cdot [K]\nabla_{b\|})T, 1\rangle_{\mathrm{GS}} = \langle T, 1\rangle_{\mathrm{GS}} \tag{3.21}$$

The proposed scheme therefore preserves three properties of the continuous operator, namely self-adjointness, positivity and conservativity.

3.2.3 Discretizations of the perpendicular gradient $\nabla_{b\perp}$ and Laplacian

In this work, we propose a conservative approach to discretize the operators in the perpendicular direction. In order to maintain the stencil size, the stencil that will be used matches that used in the parallel direction and defined in Sec. 3.2.1. The perpendicular gradient is estimated at the same points in FS, commonly shared with the surrounding control volumes (CVs). Indeed, perpendicular gradient is defined in FS from the discrete expression of its integral definition (Eq. 3.5):

$$\nabla_\perp T = \lim_{V(K)\to 0} \frac{1}{\|\mathbf{e}_\perp^1\| V(K)} \int_{S(K)} T\boldsymbol{e}_\perp^{\mathbf{1}} \cdot \boldsymbol{n} dS \tag{3.22}$$

where $V(K)$ and $S(K)$ are the volume and enclosing surface of the compact K, which leads to the discrete definition:

$$\widetilde{\nabla_{\perp} T_p} = \frac{1}{\|\mathbf{b}^1_{\perp}\| \widetilde{\Delta V_p}} \sum_q T^{int}_{qp}\, a_{qp}\, \boldsymbol{b}^1_{\perp\, qp} \cdot \boldsymbol{n}_{qp}, \tag{3.23}$$

where $\widetilde{\Delta V_p}$ is the volume defined in Sec. 3.2.1, q the number of $\widetilde{\mathrm{CV}}_p$ faces (see for example on Fig. 3.7), $\boldsymbol{b}_{\perp\, qp}$ the perpendicular vector to $\boldsymbol{b}$ (also perpendicular to the considered plane). Eq. 3.23 requires the evaluation of the flux $\widetilde{\nabla_{b\perp} T_p}$ at the barycenters of all faces. To maintain the stencil size, derivatives in y and z directions are evaluated at the points (x_i, y^-, z^-) and (x_{i+1}, y^+, z^-) in both $\mathcal{X}_i$ and $\mathcal{X}_{i+1}$ planes (Eq. 3.2).

3.3 Construction of the stencils in a 2D domain

For simplicity, the construction is presented here in 2D but the generalization to 3D problems is straigthforward although cumbersome to write. The corresponding test case is presented in Sec. 3.5.3. The diffusion lines are considered as straight and parallel, and the diffusion coefficient can be non uniform. The same notation as introduced in Sec. 3.2 are used, volumes and surfaces becoming surfaces and lengths, respectively. Since the aligned coordinates are local, the geometrical origin is established at T_{ij}.

3.3.1 Parallel Laplacian operator

Geometrical definitions

The definition of the fluxes space between adjacent control volumes is based on the geometrical definitions of the parallel diffusion lines $b_{\|}$. The local CV of $T_{i\,j}$ is bounded by the parallel field lines $b_{\|\, ij\pm 1/2}$ defined between $\mathcal{X}_{i\pm 1/2}$, Fig. 3.4. Considering the forward sense (+), the local CV is here in contact with two adjacent CV defined in the $\mathcal{X}_{i\pm 1}$ planes for the grid nodes $T_{i+1\, j+\xi}$ and $T_{i+1\, j+\xi+1}$ (note $\xi = 1$ in Fig. 3.4, see Eq. 3.47). Considering $d_{b\|}$ the distance between two adjacent $\mathcal{X}$ planes when moving along the diffusion line as:

$$d_{b\|} = \frac{\Delta x}{\cos(\alpha)}, \tag{3.24}$$

the projection of $d_{b\|}$ in the y-direction writes:

$$y_{d_{b\|}} = d_{b\|} \sin(\alpha) = \Delta x \tan(\alpha).$$

The contact surfaces (Fig. 3.4b) lead to the following areas a_1 and a_2 (here lengths) such that:

$$a_1 = \Delta y + y_{d_{b\|}} - \Delta y\,(\xi) = d_{b\|} - (\xi)\Delta y$$

$$a_2 = \Delta y - a_1,$$

The y-coordinates of the two barycenters of a_1 and a_2 express as:

$$y_{bc_1} = \frac{1}{2}(\Delta y + y_{d_{b\|}} - a_1)$$

$$y_{bc_2} = \frac{1}{2}(-\Delta y + y_{d_{b\|}} + a_2)$$

Then, the fluxes can be calculated at y_{bc_1} and y_{bc_2} in $x = (i + 1/2)\Delta x$. The control volume $\widetilde{CV}$ for each flux is limited by the parallel diffusion lines defined at $y =_{bc_1} \pm a_1/2$ and $y = y_{bc_2} \pm a_2/2$ (note $y_{bc_1} - a_1/2 = y_{bc_2} + a_2/2$, see Fig. 3.4).

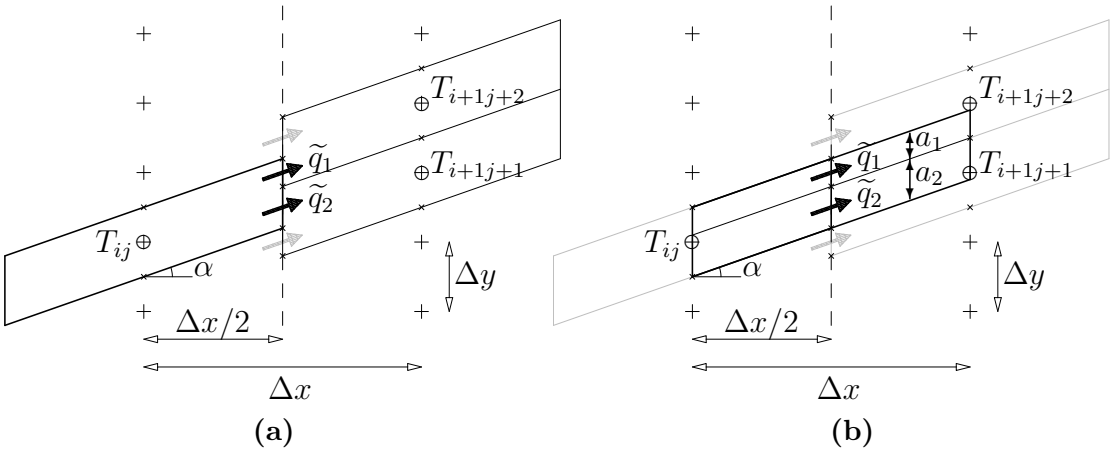


Figure 3.4: (a) 2D chart used for the flux calculation at the barycenter of the contact surface between neighboring CVs. (b) local $\widetilde{CV}$ defined by the diffusion lines surrounding the contact surfaces and bounded in the x-direction by the $\mathcal{X}_i$ and $\mathcal{X}_{i+1}$ planes.

Finite-differences step

The calculation of $\widetilde{q}_p$ reduces Eq. 3.5 to a finite-differences equation: the gradient of T at the barycenter is obtained by linear interpolation on b_{bc_p} at $\mathcal{X}_i$ and $\mathcal{X}_{i+1}$. The interpolation coordinates are obtained at x=$i\Delta x$ and $x = (i + 1)\Delta x$ as:

$$y^-_{int} = y_{bc_p} - \frac{d_{b\|}}{2},$$

$$y^+_{int} = y_{bc_p} + \frac{d_{b\|}}{2},$$

Then, a linear interpolation in the y-direction allows us to evaluate T at $y_{int}^{\pm}$. For $p = 2$, it is illustrated on Fig. 3.5:

$$\begin{aligned} T_{int\,2}^{-} &= f_{ij}^{int\,2}\,T_{ij} + f_{ij-1}^{int\,2}\,T_{ij-1} \\ &= \left[1 - \frac{1}{\Delta y}\left|\left(y_{bc_2} - \frac{y_{d_{b\|}}}{2}\right)\right|\right] T_{ij} + \frac{1}{\Delta y}\left|\left(y_{bc_2} - \frac{y_{d_{b\|}}}{2}\right)\right| T_{ij-1} \end{aligned} \tag{3.25}$$

$$\begin{aligned} T_{int\,2}^{+} &= f_{i+1j+\xi+1}^{int\,2}\,T_{i+1j+\xi} + f_{i+1j+\xi+1}^{int\,2}\,T_{i+1j+\xi+1} \\ &= \left[1 - \frac{1}{\Delta y}\left(y_{bc_2} + \frac{y_{d_{b\|}}}{2} - \xi\Delta y\right)\right] T_{i+1j+\xi} \\ &+ \frac{1}{\Delta y}\left(y_{bc_2} + \frac{y_{d_{b\|}}}{2} - \xi\Delta y\right) T_{i+1j+\xi+1} \end{aligned} \tag{3.26}$$

For $p = 1$ we get:

$$\begin{aligned} T_{int\,1}^{-} &= f_{ij}^{int\,1}\,T_{ij} + f_{ij+1}^{int\,1}\,T_{ij+1} \\ &= \left[1 - \frac{1}{\Delta y}\left(y_{bc_1} - \frac{y_{d_{b\|}}}{2}\right)\right] T_{ij} + \frac{1}{\Delta y}\left(y_{bc_1} - \frac{y_{d_{b\|}}}{2}\right) T_{ij+1} \end{aligned} \tag{3.27}$$

$$\begin{aligned} T_{int\,1}^{+} &= f_{i+1j+\xi+1}^{int\,1}\,T_{i+1j+\xi} + f_{i+1j+\xi+1}^{int\,1}\,T_{i+1j+\xi+1} \\ &= \left[1 - \frac{1}{\Delta y}\left(y_{bc_1} + \frac{y_{d_{b\|}}}{2} - \xi\Delta y\right)\right] T_{i+1j+\xi} \\ &+ \frac{1}{\Delta y}\left(y_{bc_1} + \frac{y_{d_{b\|}}}{2} - \xi\Delta y\right) T_{i+1j+\xi+1}. \end{aligned} \tag{3.28}$$

Then, the gradient calculated at bc_p writes:

$$grad(T)_p = \frac{T_{int\,p}^{+} - T_{int\,p}^{-}}{d_{b\|}}.$$

The value of $K_{b\|}$ at the barycenter is obtained here by interpolation in the same coordinates than $T_{int\,p}^{\pm}$ (Eqs. 3.25, 3.26, 3.27 and 3.28), obtaining $K_{b\|\,int\,p}^{-}$ at $x = i\Delta x$ and $K_{b\|\,int\,p}^{+}$ at $x = (i+1)\Delta x$. The aligned interpolation leads to:

$$K_{b\|\,p} = \frac{K_{b\|\,int\,p}^{-} + K_{b\|\,int\,p}^{+}}{2} \tag{3.29}$$

Then, the discrete flux is obtained as:

$$\widetilde{q}_p = K_{b\|\,p}\,grad(T)_p \tag{3.30}$$

leading to the following stencil:

$$
\begin{aligned}
\widetilde{f}_1 &= \frac{a_1}{a_1+a_2} K_{b\|\,1} f^{int\,1}_{i+1j+\xi+1} + \frac{a_2}{a_1+a_2} K_{b\|\,2} f^{int\,2}_{i+1j+\xi+1} \\
\widetilde{f}_2 &= \frac{a_1}{a_1+a_2} K_{b\|\,1} f^{int\,1}_{i+1j+\xi} + \frac{a_2}{a_1+a_2} K_{b\|\,2} f^{int\,2}_{i+1j+\xi} \\
\widetilde{f}_3 &= \frac{a_1}{a_1+a_2} K_{b\|\,1} f^{int\,1}_{ij+1} \\
\widetilde{f}_4 &= \frac{a_1}{a_1+a_2} K_{b\|\,1} f^{int\,1}_{ij} + \frac{a_2}{a_1+a_2} K_{b\|\,2} f^{int\,2}_{ij} \\
\widetilde{f}_5 &= \frac{a_2}{a_1+a_2} K_{b\|\,2} f^{int\,2}_{ij-1}
\end{aligned}
\tag{3.31}
$$

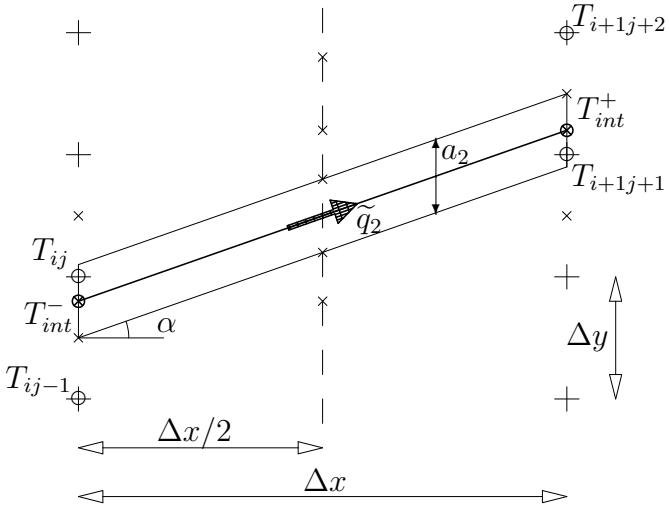


Figure 3.5: 2-D chart of the calculation of the flux $\widetilde{q}_2$ attached to $\widetilde{CV}_2$. The volume $\widetilde{CV}_2$ is here a parallelogram with $\widetilde{\Delta V}_n = \Delta x\, a_n$. T^-_{int} and T^+_{int} are the values of the field linearly interpolated from T_{ij}, T_{ij-1} and $T_{i+1\,j+1}\,T_{i+1\,j+2}$ respectively (see Eqs. 3.25, 3.26, 3.27 and 3.28).

Complete stencil of the Laplacian operator

According to Eq. 3.18, the final stencil is given by the product of the transposed sparse matrix related to the flux definition on Eq. 3.30. The stencil with the coefficients of Eqs. 3.32 is shown on Fig. 3.6 . Due to the use of SOM, the resulting discrete Laplacian matrix is positive definite.

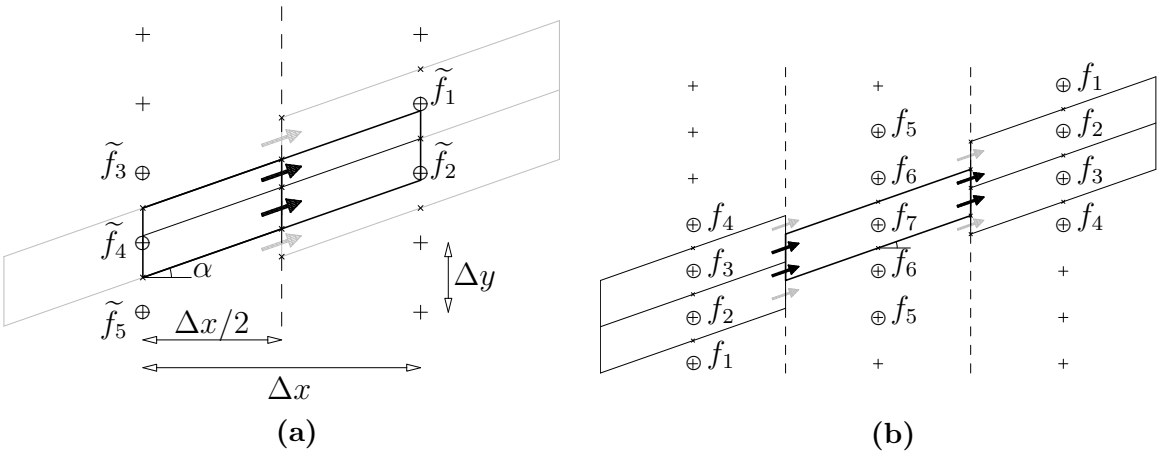


Figure 3.6: Sketches of the stencils of the discrete gradient (a) and Laplacian (b) operators.

$$
\begin{aligned}
f_1 &= -\widetilde{f}_1\widetilde{f}_5 \\
f_2 &= -\widetilde{f}_1\widetilde{f}_4 - \widetilde{f}_5\widetilde{f}_2 \\
f_3 &= -\widetilde{f}_4\widetilde{f}_2 - \widetilde{f}_1\widetilde{f}_3 \\
f_4 &= -\widetilde{f}_3\widetilde{f}_2 \\
f_5 &= -\widetilde{f}_3\widetilde{f}_5 \\
f_6 &= -\widetilde{f}_3\widetilde{f}_4 - f_5\widetilde{f}_4 - \widetilde{f}_1\widetilde{f}_2 \\
f_7 &= 1 - \widetilde{f}_4^2 - \widetilde{f}_5^2 - \widetilde{f}_1^2 - \widetilde{f}_2^2 - \widetilde{f}_3^2
\end{aligned}
\tag{3.32}
$$

The conservative discretization of the gradient (see an illustration on Fig. 3.6 together with SOM in the *present scheme* leads to a rather large stencil with 13-nodes. It can be compared to the 5-nodes (with linear interpolation) or 7-nodes (with an interpolation of degree 2) stencils of the Ottaviani's scheme and to the 7-nodes (with a linear interpolation) stencil of the Stegmeir's scheme.

3.3.2 Perpendicular gradient $\nabla_{b\perp}$ and Laplacian

Under the current assumptions, Eq. 3.23 reduces to:

$$
\widetilde{\nabla_{b\perp} T_p} = \frac{\widetilde{\nabla_y T_p} - \widetilde{\nabla_{b\parallel} T_p} \sin\alpha}{\cos\alpha}, \tag{3.33}
$$

where $\widetilde{\nabla_{b\parallel} T_p}$ is obtained using the aligned method described in Sec. 3.2.1. Considering for example $p = 2$, Fig. 3.7, the y-direction contribution is evaluated in the planes $\mathcal{X}_i$ and $\mathcal{X}_{i+1}$ as :

$$\nabla_y T_2^- \approx K_{y2}^- \frac{T_{ij+1} - T_{ij}}{\Delta y}, \tag{3.34}$$

and analogously for the gradient $\nabla_y T_2^+$, K_{y2}^- being the diffusion in the y-direction at bc_2 defined later. Finally, the gradient at the barycenter of $\widetilde{\mathrm{CV}}_p$ is obtained by interpolation along $\mathbf{b}$ as:

$$\nabla_y T_2^{int} = \frac{1}{2}(\nabla_y T_2^- + \nabla_y T_2^+), \tag{3.35}$$

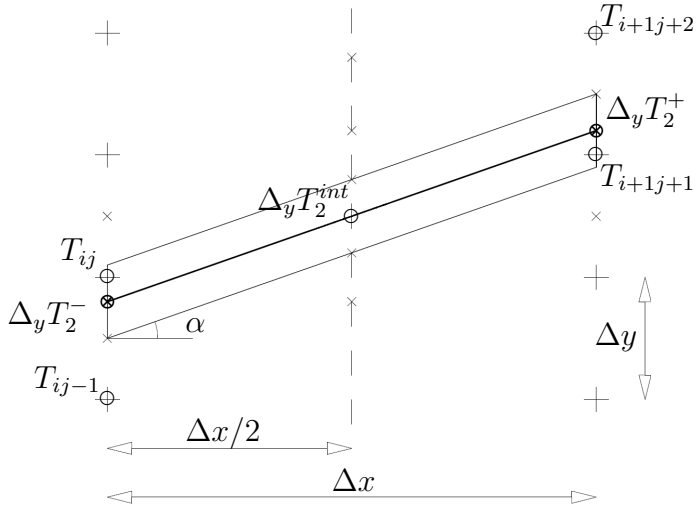


Figure 3.7: Sketch including all quantities used to obtain the gradient in the y-direction in FS for $p = 2$ (Fig. 3.2). $\nabla_y T_2^-$ and $\nabla_y T_2^+$ are evaluated on $\mathcal{X}_i$ and $\mathcal{X}_{i+1}$, respectively, and further interpolated along the parallel direction to obtain $\Delta_y T_2^{int}$ at the barycenter.

The value of the diffusion in the y-direction at bc_p is firstly evaluated at the stencil nodes, as (Fig. 3.7):

$$K_{y\,ij} = K_{b\parallel} \sin(\alpha) + K_{b\perp} \cos(\alpha)$$

Then, it is interpolated at bc_p using the stencil defined on Eq. 3.29 and based on the interpolation defined in Eqs. 3.25, 3.26, 3.27 and 3.28, Sec. 3.3.1. Once the perpendicular flux has been determined in FS, the full Laplacian is obtained by the SOM described in Sec. 3.2.2.

3.4 Numerical discretization of the boundaries

A novelty of this work is to extend to the wall the method described above by considering bounded domains. For fusion applications, cases of interest are configurations where

the flow intercepts the boundaries in the direction of anisotropy (parallel direction). For simplicity, the discretization is presented here in 2D but the generalization to 3D problems is straigthforward although cumbersome to write.

The discretization presented above must be adapted to keep the accuracy while remaining compatible with the discretization adopted for inner nodes. As a reminder, aligned methods allow the reduction of the required number of degrees of freedom in one mesh direction, by using the knowledge that the main contributions to the solutions will either be uniform or slowly varying in the main diffusion direction. One can then use a mesh with fine resolution in one direction, so as to resolve the potentially fast variations in the perpendicular direction, but with a coarse resolution in the other direction that accounts for variations in the parallel direction. The constraint to keep these benefits is that the mesh at the boundary allows to adequately represent the fast variations of the solution along the boundary itself. *Non-aligned methods* [GYKL05], or *aligned methods* using a stencil based on surrounding grid points [vEKdB14], possibly keep working near the boundaries using few ghost points (since the stencil are not oriented to $\boldsymbol{b}$). However, near the boundary, *aligned methods* for which the stencil is oriented, such as the present scheme, or in others Refs. [Ott11, HO13, SCM+16], parallel diffusion line tracking intercepts the neighbour plane outside the domain limits, Fig. 3.8a. This reason suggests another treatment of the operator near the boundary, solving Eq. 2.1 with an aligned approach and avoiding the uncertainties of far ghost points.

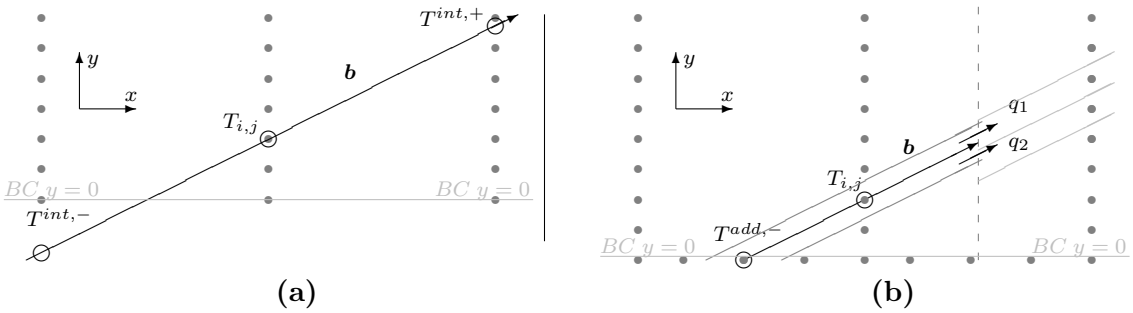


Figure 3.8: Numerical discretizations near the boundary condition $y = 0$ using *aligned methods*. (a) *Aligned method* with ghost points $T^{int,-}$ located along the parallel direction and possibly far outside the domain (extrapolated points method). (b) Present method with ghost points T^{add} added on the boundary of the domain only.

We propose here to add ghost points directly on the boundary of the domain, and located at the intersection of the parallel diffusion line with the boundary, as shown in Fig. 3.8b. Such points are needed since we may have at the two boundaries of the domain:

$$\int_{x_i}^{x_{i-1}} \frac{b_y}{b_x} dx < 0 \quad \text{at} \quad y = 0, \qquad \text{and} \qquad \int_{x_i}^{x_{i+1}} \frac{b_y}{b_x} dx > 2\pi \quad \text{at} \quad y = 2\pi, \tag{3.36}$$

depending on the resolution and the incidence α.

For any point $T_{i,j}$ located close to the grid limits at $y = 0$ and $y = 2\pi$, extra points are added in the x-direction at the coordinates:

$$x_{add} = i\Delta x + \int_0^{y_i} \frac{b_x}{b_y} dy \quad \text{at} \quad y = 0, \qquad \text{and} \qquad x_{add} = i\Delta x + \int_{y_i}^{2\pi} \frac{b_x}{b_y} dy \quad \text{at} \quad y = 2\pi. \tag{3.37}$$

These points being now located on the boundary, the value of the field may be directly obtained from the boundary conditions associated to Eq. 2.1. For Dirichlet boundary condition ($\beta = 0$ and $\gamma = 1$) the result is immediate. For Neumann boundary condition ($\beta = 1$ and $\gamma = 0$), the derivative in the parallel direction has to be evaluated using interior grid points. In this case, simplifying the label (i, j) as λ, we get:

$$(T_\lambda^{add})_{y=0} = \frac{\nabla_{b\|}(T_\lambda^{add})_{y=0}(d_{b\|\,1}^2 d_{b\|\,T} - d_{b\|\,1} d_{b\|\,T}^2) - T_\lambda d_{b\|\,T}^2 + (T_\lambda^{int})_{\mathcal{X}_{i+1}} d_{b\|\,1}^2}{d_{b\|\,T}^2 - d_{b\|\,1}^2} \tag{3.38}$$

$$(T_\lambda^{add})_{y=2\pi} = \frac{-\nabla_{b\|}(T_\lambda^{add})_{y=2\pi}(d_{b\|\,1}^2 d_{b\|\,T} - d_{b\|\,1} d_{b\|\,T}^2) + T_\lambda d_{b\|\,T}^2 - (T_\lambda^{int})_{\mathcal{X}_{i-1}} d_{b\|\,1}^2}{d_{b\|\,T}^2 - d_{b\|\,1}^2} \tag{3.39}$$

where $d_{b\|\,1}$ and $d_{b\|\,2}$ (Eq. 3.24) are the arc lengths in the parallel direction between T_λ^{add} and T_λ, and between T_λ and T_λ^{int}, and $d_{b\|\,T} = d_{b\|\,1} + d_{b\|\,2}$. Let's notice that $d_{b\|\,p}$ is the simplification of $a_p/\Delta V_p$ (Eq. 3.5) in 2D. $(T_\lambda^{int})_{\mathcal{X}_{i\pm1}}$ defines the value of the field interpolated in the plane $\mathcal{X}_{i\pm1}$.

Since the values of T_λ^{add} at $y = 0$ and $y = 2\pi$ are located along $\boldsymbol{b}$, the CV associated to T_λ (See on Fig. 3.1) is aligned with the control volume associated to T_λ^{add}. The flux discretized using finite-difference between T_λ and T_λ^{add} remains conservative. The complete operator can be calculated by considering the fluxes balance in the CV of T_λ, i.e.:

$$\nabla \cdot [K] \nabla_{b\|} T_\lambda = \frac{q_\lambda^+ - (q_\lambda^{add})_{y=0}}{\frac{1}{2}\left(d_{b\|\,1} + d_{b\|\,2}\right)}, \tag{3.40}$$

$$\nabla \cdot [K] \nabla_{b\|} T_\lambda = \frac{(q_\lambda^{add})_{y=2\pi} - q_\lambda^-}{\frac{1}{2}\left(d_{b\|\,1} + d_{b\|\,2}\right)}, \tag{3.41}$$

where $(q_\lambda^{add})_{y=0}$ and $(q_\lambda^{add})_{y=2\pi}$ are the fluxes between then inner grid point λ and the corresponding grid point added on the boundary:

$$(q_\lambda^{add})_{y=0} = \frac{T_\lambda - (T_\lambda^{add})_{y=0}}{d_{b\|\,1}}, \qquad \text{and} \qquad (q_\lambda^{add})_{y=2\pi} = \frac{(T_\lambda^{add})_{y=2\pi} - T_\lambda}{d_{b\|\,1}}, \tag{3.42}$$

and $q_\lambda^\pm$ is the total flux considering the fluxes obtained in Eq. 3.5, as follows:

$$q_\lambda^\pm = \frac{1}{A_\lambda^\pm} \sum_p K_p (\nabla_{b\|} T)_p a_p = \frac{1}{A_\lambda^\pm} \sum_p K_p a_p \sum_\mu Q_{\mu p} T_\mu \ , \tag{3.43}$$

where

$$A_\lambda^\pm = \sum_p a_p. \tag{3.44}$$

In Eq. 3.43, the matrix product $Q_{\lambda p} T_\lambda$ gives the fluxes through the CV, and $\pm$ represents the relative position into the CV associated to T_λ. Note than in Eq. 3.43 $A_\lambda^\pm$ and a_p reduce to lengths since the problem is 2D.

3.5 Numerical tests

Numerical tests have been performed on Eq. 2.1 in 2D in the (x, y)-plane. Thus, the rotation matrix $\boldsymbol{\mathcal{R}}$ of Eq. 2.2 defines as:

$$\boldsymbol{\mathcal{R}} = \begin{bmatrix} \cos\alpha & -\sin\alpha \\ \sin\alpha & \cos\alpha \end{bmatrix},$$

where the angle α measures the non-alignement between the unit vector along the anisotropy direction $\boldsymbol{b}$ and the x-axis. Thus, it results $\boldsymbol{b} = (\cos\alpha, \sin\alpha, 0)^t$ (Fig. 3.9). $\boldsymbol{b}$ will be assumed constant in the tests.

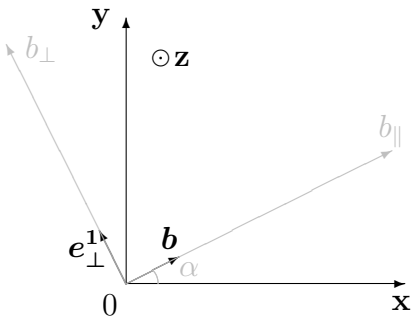


Figure 3.9: Directions of the principal axes of the diffusion tensor in the (x, y)-plane. α defines the misalignment angle of the principal axes with respect to grid points directions.

In all tests $\mu = 1$ (see a discussion in A), except for the special case of the Poisson's equation.

The following discretization methods have been considered for the tests:

- The *classical method*, refering to the asymmetric approach [vEKdB14].
- The *Günter's method*, refering to the symmetric approach proposed by Günter *et al.* [GYKL05],
- The *Ottaviani's method*, refering to an aligned approach (oriented stencil) based on a second-order parallel, polynomial interpolation [Ott11, HO13].
- The *Stegmeir's method*, refering to an aligned approach based on a linear interpolation [SCM+16].
- The *present method* refering to the work presented in this chapter. It extends *Stegmeir's* method to a conservative discretization in both parallel and perpendicular directions and to an efficient discretization of the boundary condition in bounded domains.

The first two methods used stencils independent of the diffusion tensor, and thus lie in the class of *non-aligned methods*. In contrast, the other methods lie in the class of *aligned methods*, as defined in Chapter 2.

When involved in the tests, the perpendicular part of the diffusion operator is discretized using the scheme proposed in this work in Sec. 3.2.3 for both *Stegmeir's method* and *Ottaviani's method*, which originally do not address the discretization of this flow direction.

3.5.1 Numerical details

The following manufactured source term S_a has been considered, corresponding to the superposition of a constant, an aligned and a non-aligned contribution:

$$S_a(x, y) = C_1 + C_2 \cos(m_y y + m_{x,1} x) + C_3 \sin(m_{x,2} x) \tag{3.45}$$

This source term corresponds indeed to the superposition of fluctuations varying rapidly in the perpendicular direction while being uniform along the parallel direction. The angle $\phi = tan^{-1}(m_{x,1}/m_y)$ defines the orientation of the aligned modes. The non-parallel modes vary only in x. In the case where $\phi = \alpha$, α being the pitch angle, the resolution of Eq. 2.1 with $K_{b\perp} = 0$ leads to the following solution:

$$T_a(x, y) = C_1 + C_2 \cos(m_y y + m_{x,1} x) + \frac{1}{1 + K_{b\parallel} m_{x,2}} C_3 \sin(m_{x,2} x) \tag{3.46}$$

The fluctuations related to the first term should then dominate, and the damping coefficient of this particular contribution is a good indicator of the quality of the discretization used.

All accuracy tests have been performed with $\alpha = \tan^{-1}(4/27)$, $m_y = 27$, $m_{x,1} = 4$, $m_{x,2} = 2$, $C_1 = 0$, $C_2 = 1$, $C_3 = 0.25$. 2D plots of S_a are shown on Fig. 3.10 for both a small ($K_{b\parallel} = K_{b\perp}$) and a large ($K_{b\parallel} = 10^6 K_{b\perp}$) anisotropy, showing or not parallel modulations in the parallel direction, respectively.

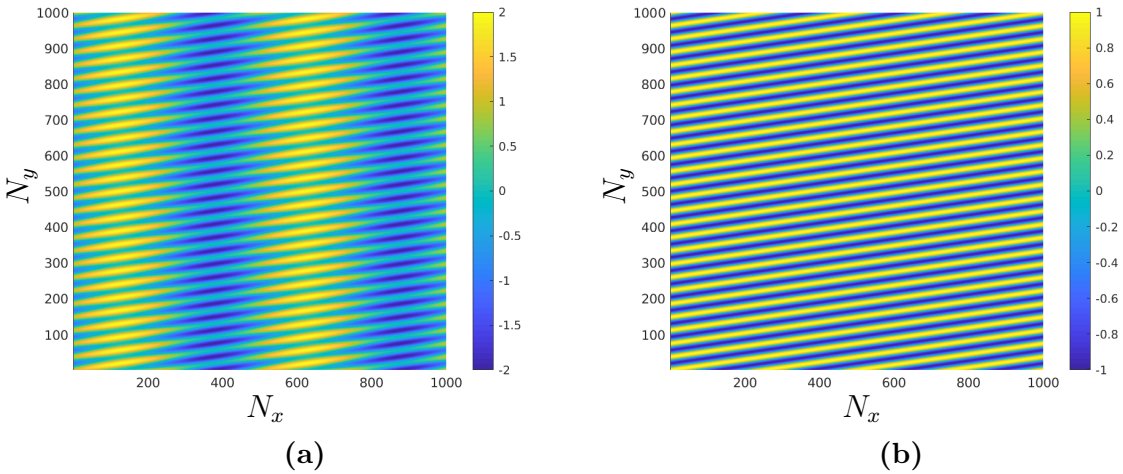


Figure 3.10: 2D plots of T_a when solving Eq. 2.1 analytically with source term S_a Eq. 3.45 for (a) an isotropic case $K_{b\parallel} = K_{b\perp}$ and (b) a strongly anisotropic case $K_{b\parallel} = 10^6 K_{b\perp}$. The source term S_a has following parameter values: $C_1 = 0$, $C_2 = 1$, $C_3 = 1$, $m_y = 27$, $m_{x,1} = 4$ and $m_{x,2} = 2$.

Tests are made by keeping fixed the resolutions in the x-direction ($N_x = 8$, 16 and 32) while varying N_y such that $(N_{dof})_{max} = \max(N_x \cdot N_y) = 512^2$.

3.5.2 Error estimate

Numerical tests in Appendix B show that the $\mathcal{H}^1$-error (described in Appendix C) is better suited than the classically used $\mathcal{L}^2$-error to evaluate the accuracy of these schemes. Indeed, the $\mathcal{L}^2$-error eventually leads to misleading behaviour due to some eventual spurious aliasing effect.

For all tests, only the optimal values of the error are retained, as illustrated in Fig. 3.11a. They correspond to the minimal N_{dof} and $\mathcal{H}^1$-error relation. For each fixed resolution in the x-direction, the error is foremost dominated by the interpolation error in the y-direction for the discretization of the parallel Laplacian of the aligned fluctuations and decreases when N_y increases. Let's notice that for a given value of N_{dof} the smallest error is obtained with the smallest resolution in the x-direction since it is associated to the largest resolution in the y-direction.

From a certain value of N_{dof} (which depends on N_x), the error stops decreasing and becomes dominated by the error made in discretizing the parallel Laplacian of non-

aligned fluctuations of the solution. N_x being fixed implies that the parallel step-size is constant, and the error, therefore, converges to a constant value.

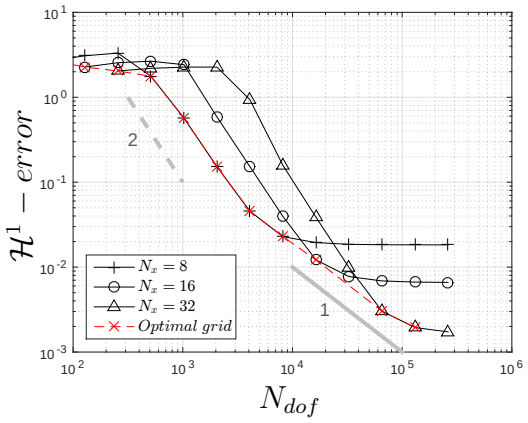


Figure 3.11: (a) Plots of the $\mathcal{H}^1$-error obtained for the *present method* when increasing N_{dof} and for three resolutions in the x-direction. For each value of the error, only the point corresponding to the lowest resolution is kept (red dotted line).

3.5.3 Accuracy tests in a 2D periodic domain

The Eq.2.1 with $\mu = 1$ and periodic boundary conditions in x and y direction is considered.

Accuracy tests for a non-zero $K_{b\perp}$

Convergence results are presented in Fig. 3.12 for an isotropic ($K_{b\parallel} = K_{b\perp}$) and an anisotropic ($K_{b\parallel} = 10^6 K_{b\perp}$) diffusion tensor (Eq. 2.21). When the tensor is isotropic, there is no significant difference between all the methods, and the errors converge at nearly the same rate, Fig. 3.12a. However, when the tensor becomes anisotropic, Fig. 3.12b clearly shows the superiority of the *aligned methods*, owing to their better ability to capture the uniformity of the dominant contribution along the vector $\mathbf{b}$. As expected by construction, the *present method* behaves similarly in this case as the two other *aligned methods* of Ottaviani and Stegmeir. However, the *classical method* fails to converge, and though it converges, the *Günter's method* requires many more points for a given accuracy.

Accuracy tests for $K_{b\perp} = 0$

We now focus on the parallel flux estimate, which is the largest source of error in such computation, and we assume that $K_{b\perp} = 0$.

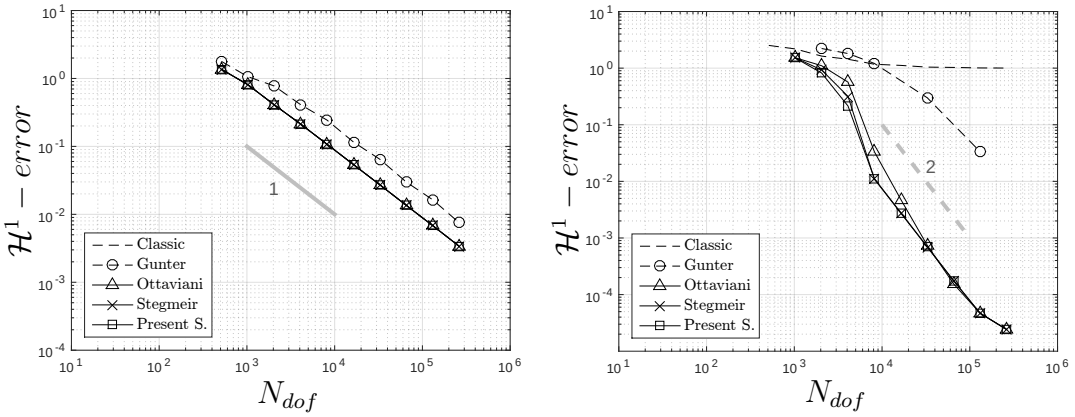


Figure 3.12: $\mathcal{H}^1$-error convergence for an isotropic ($K_{b\|} = K_{b_\perp}$) (a) and an anisotropic ($K_{b\|} = 10^6 K_{b\perp}$) (b) diffusion tensor. Bi-periodic computational domain.

Convergence results obtained for all methods are presented in Fig. 3.13 and for two values of the parallel diffusion, $K_{b\|} = 1$ and 10^6. These values correspond to solutions where parallel fluctuations are weakly or strongly damped, respectively.

For $K_{b\|} = 1$ (Fig. 3.13a), the *present method* behaves as the other two *aligned methods* of the literature, and leads to a much better convergence rate than *non-aligned methods*. In addition, the three *aligned methods* need fewer points for a given accuracy, illustrating their superior ability to accurately compute the parallel Laplacian. For an error of about 10^{-2}, they indeed need about 10 times fewer points. The shift in the convergence rate between 2 and 1 at $N_{dof} = 4096$, already observed in Fig.3.11, corresponds to change in the structure of the error. For $N_{dof} \leqslant 4096$, the error is dominated by the interpolation error in the y-direction, required by all *aligned methods* to evaluate the parallel gradient. This error decreases when N_y increases, the resolution in the x-direction being fixed. For $N_{dof} > 4096$, the error does no longer depend on the resolution in the y-direction but only on the resolution in the x-direction.

For $K_{b\|} = 10^6$ (Fig. 3.13b), the *non-aligned methods* fail to converge regardless of the resolution. In this case, the *aligned methods*' trend is fully related with the interpolation method, the fluctuations with parallel variations being strongly damped (the third component of Eq. 3.45 vanishes), the problem becomes a fully aligned problem constant in the parallel direction. The *present method* provides the best result on this test. The difference with the *Stegmeir's method* is small whatever the resolution, but it is larger with the *Ottaviani's method*, particularly at moderate resolutions. The difference with this latter decreases at high resolutions.

For a practical point of view and codes users, it is greatly useful to determine which is the resolution needed depending on the targeted accuracy, the parallel diffusion and the chosen discretization method. We show 2D plots of the $\mathcal{H}^1$-error as a function of N_x

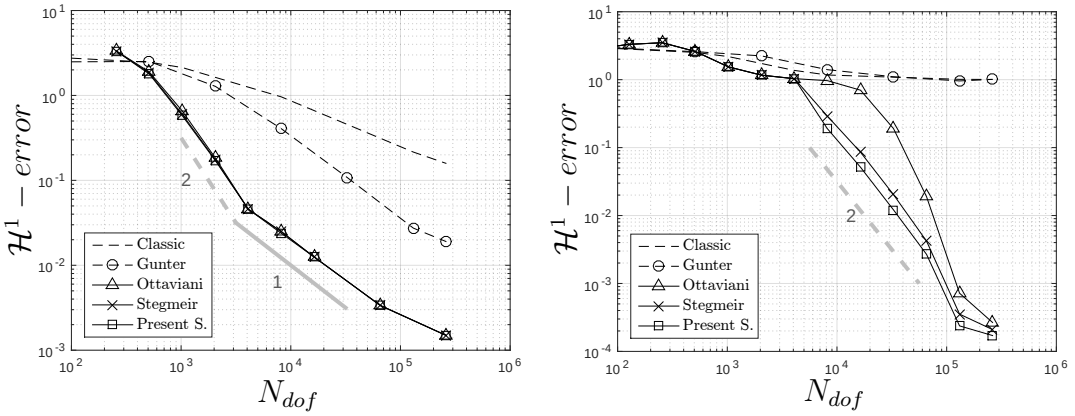


Figure 3.13: $\mathcal{H}^1$-error convergence in a bi-periodic computational domain with $K_{b\perp} = 0$:(a) $K_{b\parallel} = 1$ and (b) $K_{b\parallel} = 10^6$.

and N_y and for $K_{b\parallel} = 1$ and $K_{b\parallel} = 10^6$ that corresponds to solutions for which parallel fluctuations are weakly or strongly damped, respectively.

For the two *non-aligned methods*, the *Classic* and the *Günter's* methods, 2D plots are shown on Figs. 3.14, 3.15, respectively. For low parallel diffusion, $K_{b\parallel} = 1$, results show that the converge requires a minimal resolution in the x-direction ($N_x \approx 16 - 32$). However, for large parallel diffusion, $K_{b\parallel} = 10^6$, the *Classic* method does not converge to the solution for any tested resolution. For the *Günter's* method, the convergence domain is very small, and the method converges only at high resolutions, from 128×1024. Then, not aligned methods requires very large resolutions in both $x - y$ directions when solving highly anisotropic cases.

For the three *aligned methods*, i.e. the *Ottaviani's, Stegmeir's* and the *present* methods, results are shown in Figs. 3.16, 3.17, 3.18, respectively. As soon as the resolution in the y-direction is sufficient, the methods converge with only a few points in the x-direction. At low parallel diffusion $K_{b\parallel} = 1$, the accuracy is only weakly sensitive to N_x and a large improvement in accuracy can be gained by increasing only N_y. At large parallel diffusion $K_{b\parallel} = 10^6$, results show the superiority of the aligned methods with respect to non-aligned methods. The numerical diffusion is reduced here with N_x, showing a rapid convergence from $N_y \geq 512$. At high resolutions in the x-direction, the aligned methods do no longer converge due to the interpolation error in the y-direction. The latter has a bigger impact on finite differences when Δx (which is in the denominator of finite difference) becomes small. This effect is amplified otherwise by the value of $K_{b\parallel}$.

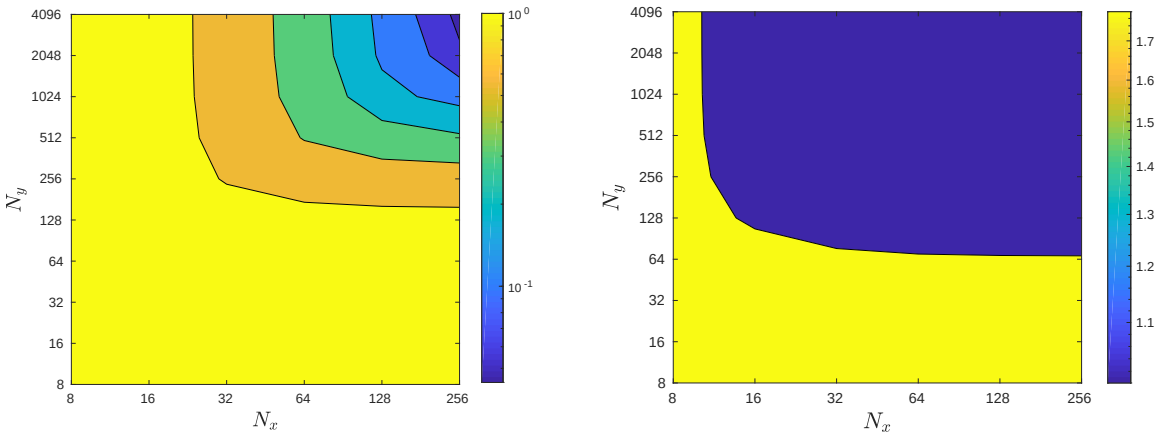


Figure 3.14: 2D plots of the $\mathcal{H}^1$-error as a function of N_x, and N_y for the *Classic method* and $K_{b\perp} = 0$. (a) $K_{b\parallel} = 1$, and (b) $K_{b\parallel} = 10^6$. The computational domain is bi-periodic.

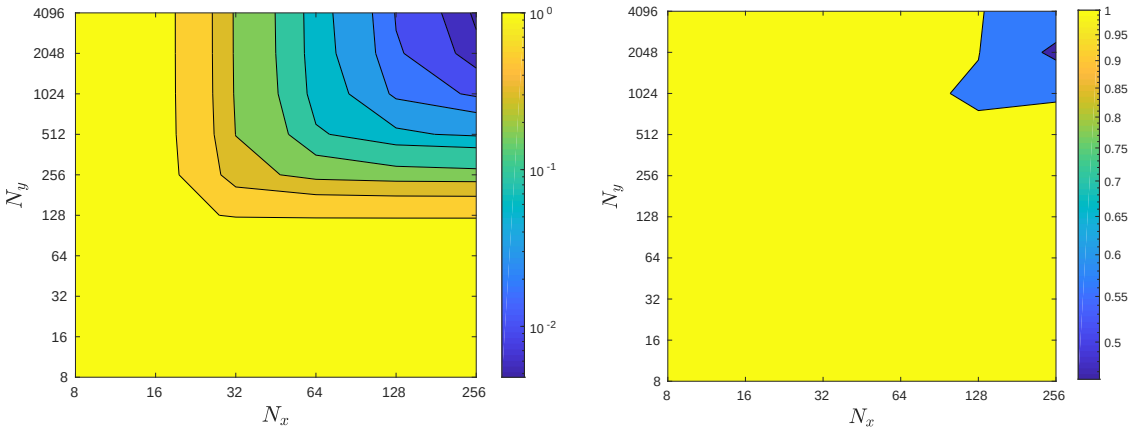
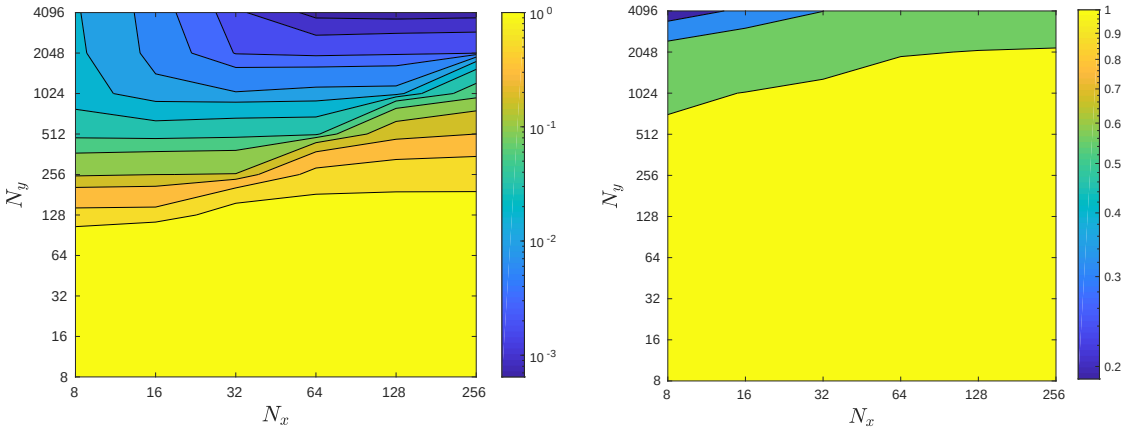


Figure 3.15: 2D plots of the $\mathcal{H}^1$-error as a function of N_x, and N_y for the *Günter's method* and $K_{b\perp} = 0$. (a) $K_{b\parallel} = 1$, and (b) $K_{b\parallel} = 10^6$. The computational domain is bi-periodic.

Tests of conservativity

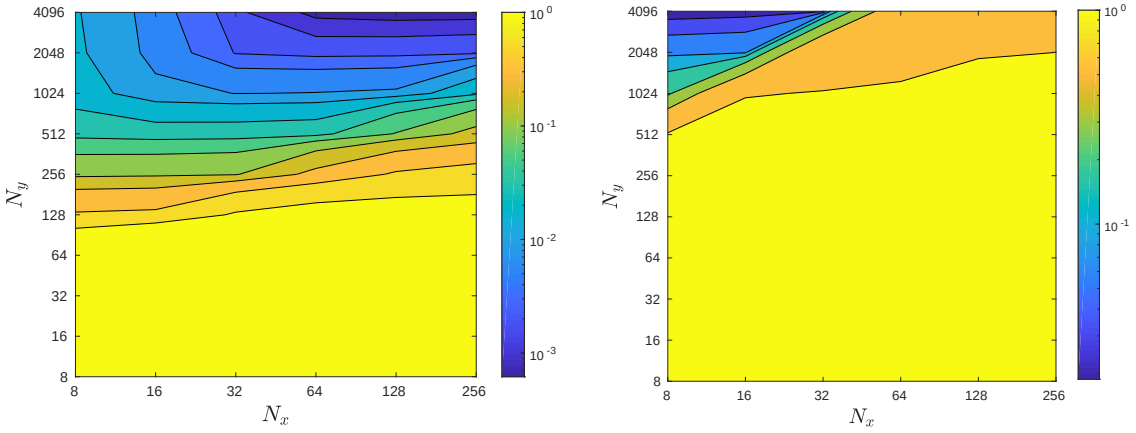
A new feature of the *present method* with respect to existing *aligned methods* of the literature is to involve a conservative discretization of the fluxes. It is shown here for the parallel operator, the discretization of the perpendicular operator (Sec. 3.2.3) implicitly guaranteeing the conservativity in this direction.

In *aligned methods* in the literature, *Ottaviani's* and *Stegmeir's methods* evaluate fluxes at the center of the CVs faces for each plane $\mathcal{X}_i$. This leads to a misalignement of the fluxes between adjacent CVs (Fig. 3.19a) and therefore to a non-conserving scheme.



(a) **(b)**

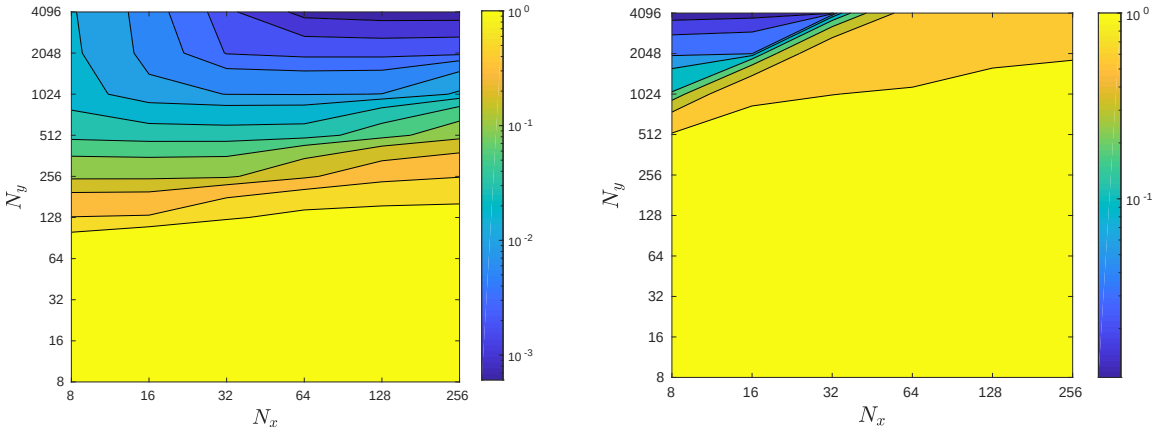
Figure 3.16: 2D plots of the $\mathcal{H}^1$-error as a function of N_x, and N_y for the *Ottaviani's method* and $K_{b\perp} = 0$. (a) $K_{b\parallel} = 1$, and (b) $K_{b\parallel} = 10^6$. The computational domain is bi-periodic.



(a) **(b)**

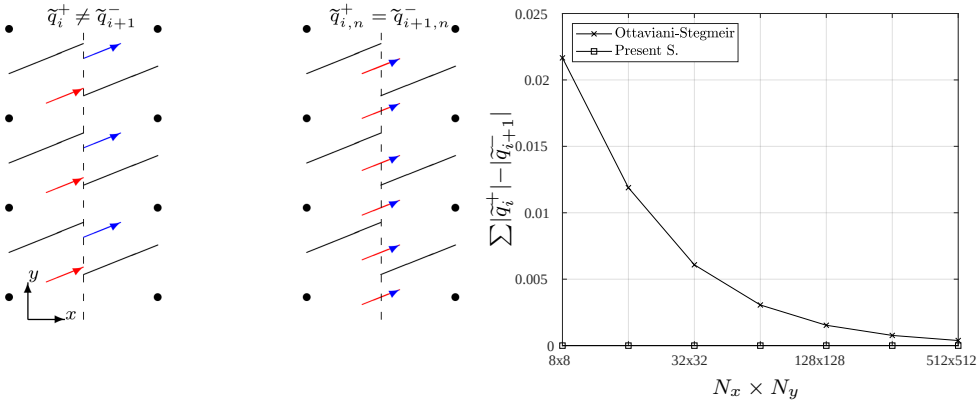
Figure 3.17: 2D plots of the $\mathcal{H}^1$-error as a function of N_x, and N_y for the *Stegmeir's method* and $K_{b\perp} = 0$. (a) $K_{b\parallel} = 1$, and (b) $K_{b\parallel} = 10^6$. The computational domain is bi-periodic.

The discretization of the fluxes calculated at the center of the common faces of two adjacent CVs ensures the conservativity of the *present method*, Fig. 3.19b. This flux definition leads to symmetric definition of fluxes between $\mathcal{X}$ planes independently of $K_{b\parallel}$. To show that, a test has been carried out considering the source term $T_a = 2 + \sin(x)\sin(y)$ with a unhomogeneous $K_{b\parallel} = 2 + \sin(x)\sin(y)$. The test considers different N_{dof}, showing the quantity $|\widetilde{q}_i^+| - |\widetilde{q}_{i+1}^-|$, representing the balance between the red and blue fluxes represented in Fig. 3.19a for *Ottaviani's* and *Stegmeir's methods* (both schemes lead the same flux defintion here; see [SCM+16]) and Fig. 3.19b for the Present S. The



(a) **(b)**

Figure 3.18: 2D plots of the $\mathcal{H}^1$-error as a function of N_x, and N_y for the *present method* and $K_{b\perp} = 0$. (a) $K_{b\|} = 1$, and (b) $K_{b\|} = 10^6$. The computational domain is bi-periodic.



(a) **(b)** **(c)**

Figure 3.19: Sketches showing flux discretizations between adjacent control volumes for *Ottaviani's method* (a) aligned to the grid points [Ott11] and (b) the *present method* centered to the contact surface between control volumes. (c) Plot for different grids $N_x \times N_y$ of the relative difference $\sum|\tilde{q}_i^+|-|\tilde{q}_{i+1}^-|$ between the forward fluxes of $\mathcal{X}_i$ plane (red fluxes in sketch (a)) and the backward fluxes of $\mathcal{X}_{i+1}$ plane (blue fluxes in sketch (a)).

test results, Fig. 3.19c, show how the symmetric and unique definition of the *Present S.* fluxes leads to a perfect flux balance equal to zero, which means the difference of the quantity $|\tilde{q}_i^+|-|\tilde{q}_{i+1}^-|$ is always equal zero by construction. For *Ottaviani-Stegmeir* $|\tilde{q}_i^+|-|\tilde{q}_{i+1}^-|$ the non symmetric definition of fluxes lead to an unbalanced sum, which rises for lower N_{dof}.

3.5.4 Accuracy tests in a 2D bounded domain

Another new feature presented here is an efficient and accurate discretization of the boundary conditions in bounded domains for aligned methods, which are involved in many realistic applications although they have been much less investigated in the relevant literature.

Eq.2.1 with $\mu = 1$ and Dirichlet boundary conditions is considered. In all tests, the perpendicular diffusion $K_{b\perp} = 0$.

Three discretizations of the boundary condition have been formerly mentioned in Sec. 3.4: added points aligned along b (extrapolated grid points in the y-direction, Fig. 3.20a), the *Günter method* which relies solely on points already in the grid, and finally added points on the boundary (added points), which is the new discretization proposed in this thesis (Fig. 3.20b).

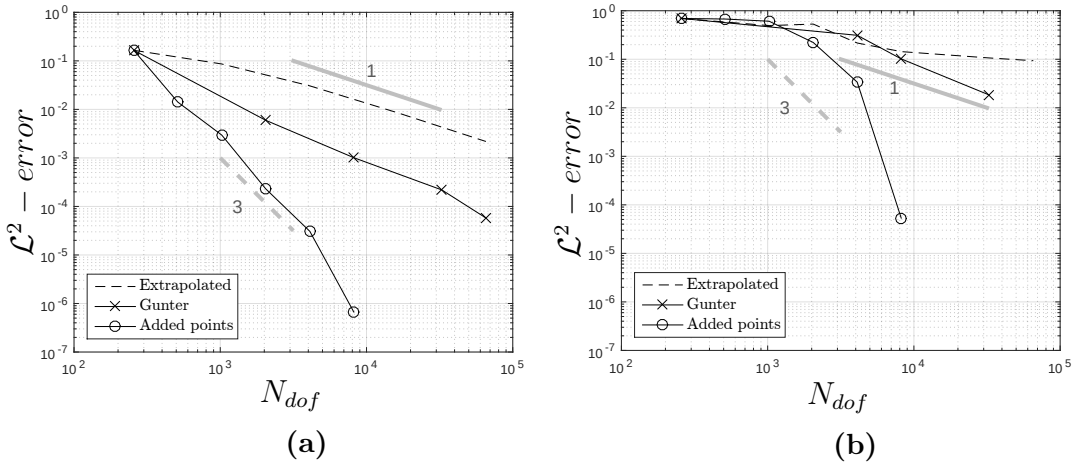


Figure 3.20: $\mathcal{L}^2$-error convergence of the *present method* for $K_{b\parallel} = 1$ (a) and $K_{b\parallel} = 10^6$ (b) and for three discretizations of the Dirichlet boundary condition. $K_{b\perp} = 0$.

On Fig. 3.20 the *present method* is tested with the three discretizations of the boundary condition. The $\mathcal{L}^2$-error has been retained as the blending of aligned methods for interior points and non-aligned methods in the neighbourhood of the boundaries, which use different discretizations of gradients, makes the evaluation of the $\mathcal{H}^1$-error problematic. Furthermore, the $\mathcal{L}^2$-error is sufficient here to qualify the differences in accuracy between the proposed boundary discretizations. Extrapolated grid points in the y-direction show poor performances, in particular for $K_{b\parallel} = 10^6$, since the rapid variations in the parallel direction limit the accuracy of the extrapolation, as explained in Sec. 3.4. The use of a non-aligned approach like with the *Günter's method* in the discretization of the boundary condition needs only one ghost point in the y-direction, but the ratio N_x/N_y is out of the limit of Nyquist-Shannon theorem provided for *non-aligned methods* when *aligned methods* reach higher performance. The added points discretization proposed in

this work maintains the convergence found in bi-periodic cases even if it slightly increases the number of global unknowns required. Indeed, the number of added points is equal to $2N_x\xi$ ($\ll N_x \times N_y$), where ξ defines the shift of the grid as:

$$\xi = \lfloor \frac{\Delta x}{\Delta y} \tan \alpha \rfloor, \tag{3.47}$$

considering $\mathbf{b}$ and the diffusion tensor [K] as uniform in Ω.

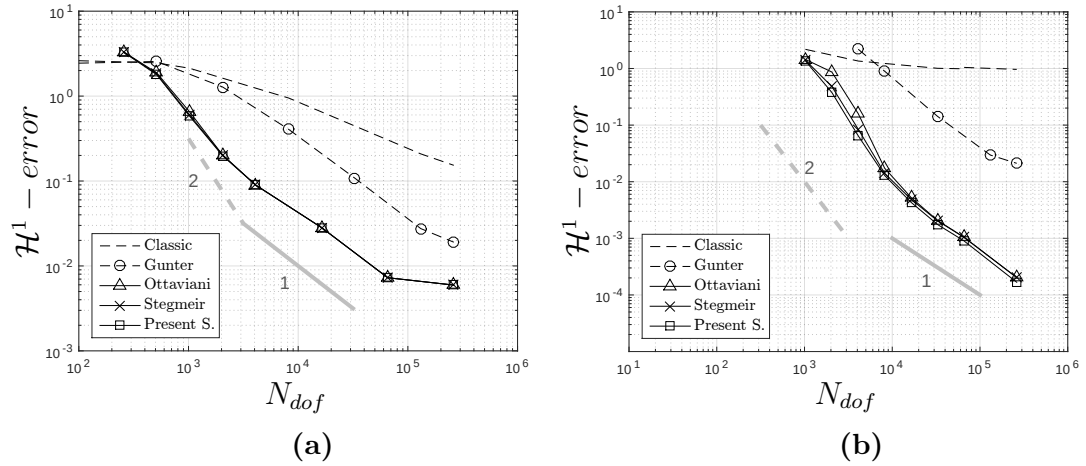


Figure 3.21: $\mathcal{H}^1$-error convergence with Dirichlet boundary conditions for all methods. For all *aligned methods* the added points discretization is used. (a) $K_{b\|} = 1$ and (b) $K_{b\|} = 10^6$.

On Fig. 3.21, accuracy tests are presented for all methods. The added points discretization proposed here is used for the three *aligned methods* (Ottaviani's, Stegmeir's and present). This new discretization of the boundary works well whatever the *aligned method* used in the interior of the domain. It allows recovering the good general trends obtained for the bi-periodic configuration (Fig. 3.13). As previously, the *present method* associated with this new discretization of the boundary provides the best results.

The Poisson equation

A special case of prime importance in many physical models (for example in the search of stationary solutions to the heat equation) is the case of the Poisson's equation. We consider here the general case where it is associated to Robin boundary conditions such that:

$$-\boldsymbol{\nabla} \cdot (\boldsymbol{\mathcal{K}} \cdot \boldsymbol{\nabla}))T = S \qquad \text{on } \Omega \tag{3.48}$$

$$\frac{1}{R}\nabla_{b\|}T + T = s \qquad \text{in } \Gamma \tag{3.49}$$

All values of R lead to regular problems. However, if the limit $R \to +\infty$ is very well-behaved, as one then approaches a problem with Dirichlet boundary conditions, the limit $R \to 0$ can be demanding, as one then approaches a problem with Neumann boundary conditions which are known to be-ill posed. Robin boundary conditions with $R = 1$ and $R = 10^{-3}$ are tested here.

For $R = 1$, the weight of the Dirichlet and the Neuman part in the Robin boundary condition is the same. Fig. 3.22 shows that *aligned methods* associated to the added points approach proposed in this paper for the discretization of the boundary condition confirm their superiority for both low ($K_{b\|} = 1$) and large ($K_{b\|} = 10^6$) parallel diffusion. As previously, the *present method* tends to provide the best results, even if the differences with the *Stegmeir's* and the *Ottaviani's method* are small in this case. For both values of the diffusion, the *classical method* does not converge. *Günter's method* tends to behave slightly better than for the tests carried out in the periodic domain, and the *classical method* and *Günter's method* seem to give errors independent of the level of anisotropy.

For $R = 10^{-3}$, the Neumann part becomes dominant over the Dirichlet part in the Robin boundary conditions. As mentioned above, the resolution of the Poisson's equation becomes much more demanding. This appears in the results shown in Fig. 3.23. The *classical method* does not converge whatever the parallel diffusion (as for the case $R = 1$), and, if *Günter's method* continues to converge, its convergence rate is strongly reduced. *Aligned methods* visibly still fare better than *non-aligned methods*. Surprisingly, increasing the anisotropy of the diffusion tensor improves their efficiency. The *present method* provides here similar results to *Stegmeir's method*.

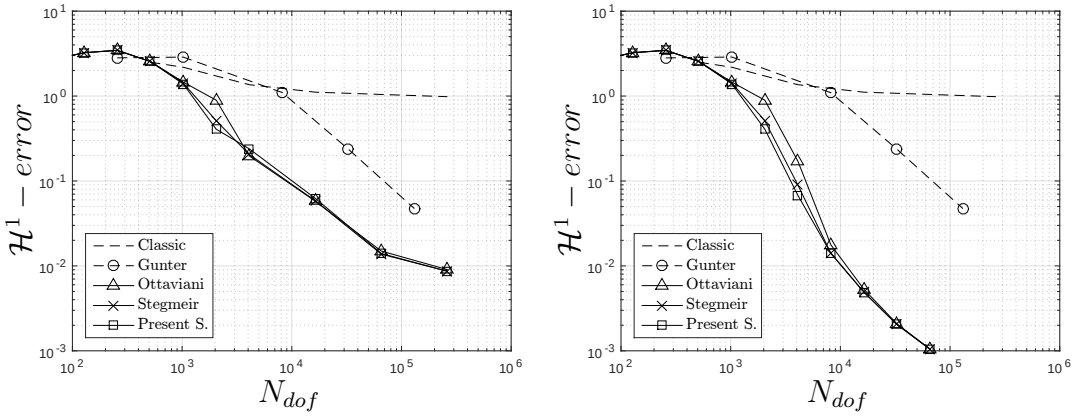


Figure 3.22: $\mathcal{H}^1$-error convergence for the Poisson's equation Eq. 3.48 with Robin boundary condition with $R = 1$. $K_{b\|} = 1$ (a) and $K_{b\|} = 10^6$ (b). $K_{b\perp} = 0$.

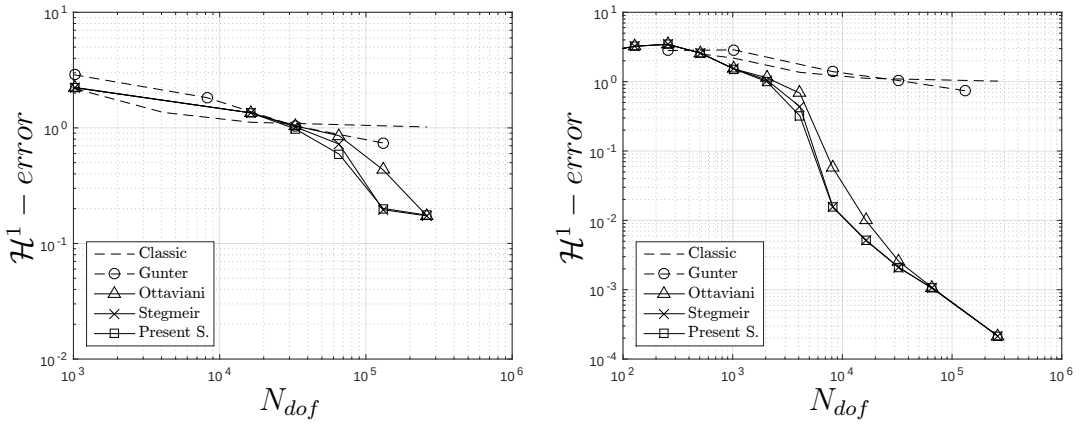


Figure 3.23: $\mathcal{H}^1$-error convergence for the Poisson's equation Eq. 3.48 with Robin boundary condition with $R = 10^{-3}$. $K_{b\parallel} = 1$ (a) and $K_{b\parallel} = 10^6$ (b). $K_{b\perp} = 0$.

Chapter 4

Invariant field decomposition

4.1 Introduction

In the previous chapter, we have shown that when solving an anisotropic diffusion problem, most of the spectral content of the solution consists in modes that are completely or nearly aligned along the anisotropy diffusion direction, since non-aligned modes are strongly damped. These aligned modes with rapid field variation in the y-direction are the most impacted by numerical diffusion.

Even though aligned methods are well adapted to such modal representation, their implementation involves the reconstruction of the solution using interpolations in the y-direction. This direction typically involves short-wavelength variations in the solution, and so this reconstruction can largely contribute to the error. In the previous chapter, we have shown that such a problem can be handled by defining grids such that $N_y \gg N_x$, avoiding interpolation degrees higher than 1 (linear interpolation). Indeed, the use of interpolations of high degrees has several drawbacks and has to be avoided. We can mention:

- The large stencils required by these interpolations, induce more dense matrices and so longer computational times to invert the discrete diffusion matrix. Moreover, such polynomial interpolations (with degree ≥ 2) introduce matrix coefficients of different signs, that can lead to ill-conditioned matrices in highly anisotropic cases.

- Gibbs oscillations near strong gradients [BBC+13], which are commonly observed in the interpolation direction when using aligned methods, are also a source of error. Non-linear interpolation methods like PPH [ADLT06, Ama08, AL05] are specially designed to avoid such Gibbs oscillations but however, they require a previous evaluation of the field, that forces to rewrite (and re-inverse) the discrete diffusion matrix at each time step.

All these drawbacks limit the improvement of aligned methods to deal with anisotropic diffusion operators under grid misalignment conditions.

The aligned modes are dominant in the solution, and discretization and inversion errors on these modes are the most critical to the the accuracy of the solution. However, they are specific in that they are invariant when the parallel diffusion operator is applied to them (with $\mu = 1$). This chapter proposes a way of using this property to improve the accuracy and the computational efficiency of the discrete elliptic operator. The idea is to split the field T into two terms: one invariant with respect to the parallel diffusion operator, defined as $\overline{T}$, and the other one non-invariant, defined as $\widetilde{T}$, such as:

$$T = \overline{T} + \widetilde{T} \tag{4.1}$$

so that Eq. 2.1 leads to:

$$T - \boldsymbol{\nabla} \cdot (K_{b\|} \cdot \boldsymbol{\nabla})(\overline{T} + \widetilde{T}) = T - 0 - \boldsymbol{\nabla} \cdot (K_{b_\|} \cdot \boldsymbol{\nabla})\widetilde{T} = S_a \tag{4.2}$$

4.1.1 Mathematical model

In order to compare with the results of the previous chapter, the Helmholtz equation is considered here for $\mu = 1$ and with bi-periodic boundary conditions such as:

$$T - \boldsymbol{\nabla} \cdot (\boldsymbol{\mathcal{K}}_{b\|} \cdot \boldsymbol{\nabla})T = S_a, \;\; in\ \Omega \subset \mathbb{R}^3 \tag{4.3}$$

with $\Omega = [0, 2\pi] \times [0, 2\pi] \times [0, 2\pi]$. Aligned fluctuations belong to the null set of the parallel diffusion operator. Using this, the space of solutions can be split into two subspaces, $V_{b\|} = \{f(x, y, z) \in \mathbb{R}^3 / \nabla_{b\|} f = 0, \quad \forall (x, y, z) \in \Omega\}$ and its complementary denoted $V_{b\perp}$. Denoting $P_{b\|}$ the projector onto $V_{b\|}$, and $D_{b\|}$ the discrete analog of $\boldsymbol{\nabla} \cdot (\boldsymbol{\mathcal{K}}_{b\|} \cdot \boldsymbol{\nabla})$, one can split the original diffusion problem into two subproblems:

$$\overline{T} = P_{b\|} S_a \tag{4.4}$$

$$(1 - D_{b\|})\widetilde{T} = (1 - P_{b\|})S_a \tag{4.5}$$

yielding the solution $T = \overline{T} + \widetilde{T}$ to be solution of the problem $(1 - D_{b\|})T = S$. This approach uses a decomposition which resembles this used in the micro-macro decomposition proposed by [DLNN12]. Their approach leads to recasting the original problem as a regular saddle point problem, with better conditioning than the original problem. This recasting was done in the aim of improving the accuracy of the solution and possibly, computing time, for very large parallel diffusivities, when using iterative methods. However, let's keep the original problem Eq. 4.3 to highlight further advantages of this approach: once the discretization is performed, it allows to use the property that parallel diffusion does not affect the projection of the solution on $V_{b\|}$, and therefore it allows to eliminate on aligned fluctuations the spurious perpendicular diffusion stemming from

discretization errors, which systematically leads to artificial excessive damping. Indeed, one then observes that in the limit $K_{b\|} \to \infty$, this formulation allows to recover directly the desired properties, i.e. $\|\widetilde{T}\| = \mathcal{O}(K_{b\|}^{-1})$ and $\|\overline{T}\| = \mathcal{O}(1)$, and regardless of potential errors of discretization of the parallel diffusion.

The problem now resides in identifying a suitable projector $P_{b\|}$. We propose several projectors in the following. According to their numerical cost, we also propose several filters that, in appropriate limits, will converge to the desired projector. We particularly evaluate the impact of using filters on the accuracy of the solutions.

In Eq. 4.1, $\overline{T}$ is composed by aligned modes which, owing to the high anisotropy assumption, can have fast variations in the interpolation direction. $\widetilde{T}$, the non aligned part of T, is characterized by moderate perpendicular fluctuations (compared with $\overline{T}$), and therefore by a reduced numerical diffusion related to the numerical discretization. Numerically, considering the discrete parallel diffusion $D_{b\|}$ in Eq. 4.3, Eq. 4.2 becomes:

$$(1 - D_{b\|})(\overline{T} + \widetilde{T}) = S_a \tag{4.6}$$

If we were to know $\overline{T}$ (the invariant part), the non-invariant part could be expressed as:

$$\widetilde{T} = \frac{S_a - (1 - D_{b\|})\overline{T}}{1 - D_{b\|}} = \frac{S_a}{1 - D_{b\|}} - \overline{T}, \tag{4.7}$$

but in practice, $\overline{T}$ is part of the solution. However, if we consider now a fully aligned solution ($T = \overline{T}$) Eq. 4.2 writes:

$$\overline{T} - \boldsymbol{\nabla} \cdot (K_{b_\|} \cdot \boldsymbol{\nabla})\overline{T} = S'_a, \tag{4.8}$$

Knowing that $\boldsymbol{\nabla} \cdot (K_{b\|} \cdot \boldsymbol{\nabla})\overline{T} = 0$, the equality 4.8 leads to $\overline{T} = S'_a$: the source is also left invariant by the application of the parallel diffusion operator, and therefore a part of the solution $S'_a = \overline{S}_a$. The invariant part is already the invariant part of the right-hand side: we can split S_a, which is known, as in Eq. 4.1. Then the Eq. 4.7 becomes:

$$\widetilde{T} = \frac{S_a - (1 - D_{b\|})\overline{T}}{1 - D_{b\|}} = \frac{S_a}{1 - D_{b\|}} - \overline{S}_a, \tag{4.9}$$

finding an equivalent solution to the Helmholtz equation:

$$T = \frac{1}{1 - D_{b\|}}\widetilde{S_a} + \frac{1 - D_{b\|}}{1 - D_{b\|}}\overline{S}_a = \frac{1}{1 - D_{b\|}}\widetilde{S_a} + \overline{S}_a \tag{4.10}$$

In the following Sec. 4.2, several numerical methods are proposed to find $\overline{S}_a$. They are discussed according to their applicability depending on boundary conditions and/or implementation conditions. The effectiveness of each method is compared in Sec. 4.3, and the new problem Eq. 4.10 with no spurious perpendicular diffusion is shown in Sec. 4.4 for all discrete Laplacians described in Chapters 2 and 3.

4.2 Projection and filtering methods

4.2.1 Filtering in modal space

The first projection method is based on the Fast Fourier Transform (FFT) of T. As described in Sec. B.2 of Appendix B, T is defined in the pseudo-modal space as follows:

$$S_a(x,y) = \sum_{m,n} \hat{S}_{am,n} e^{i(nx+my)} \tag{4.11}$$

In modal space, Fig. 4.1(a) shows the source term S_a both in the physical and modal space. The plots show clear evidence of the influence of the anisotropy on the distribution of the modes (see Appendix B.2). The aligned modes can be filtered, setting the minimal bandwidth to those points aligned along the parallel diffusion direction. An aligned filter along the parallel diffusion direction is:

$$F(m,n) = \begin{cases} 1, & \text{if } m = n\,tan^{-1}(\alpha). \\ 0, & \text{otherwise.} \end{cases} \tag{4.12}$$

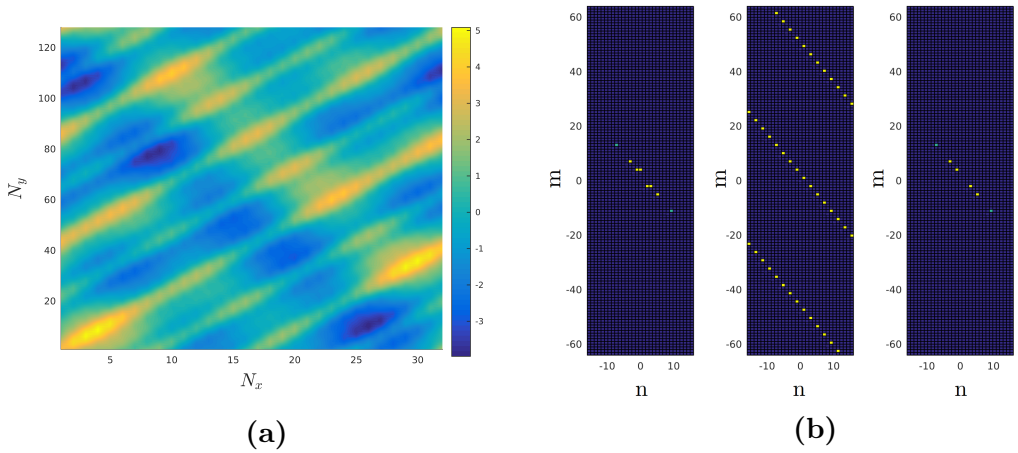
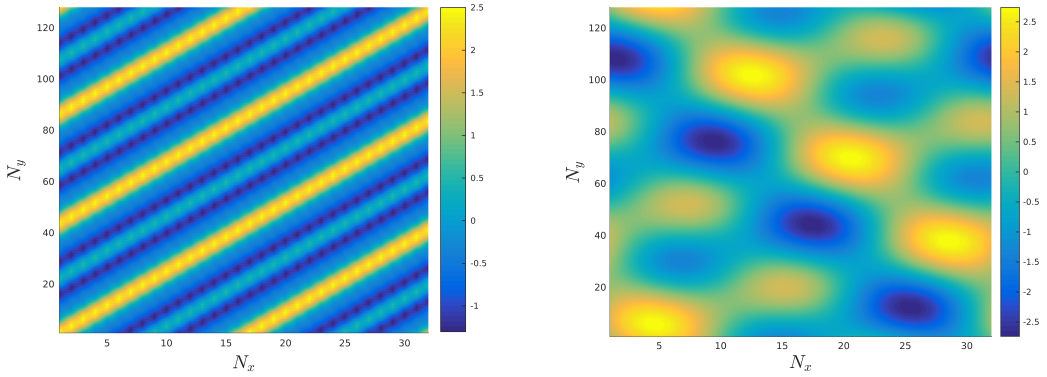


Figure 4.1: 2D plots of S_a and the filter (a) S_a in the physical space, defined by Eq. 4.25. (b) $\hat{S}_{a,m\,n}$ in the Fourier space (left), filter $F_{m\,n}$ (center), and $\hat{\overline{S}}_{mn}$ (right).

The result of the matrix product of F and S_a leads to the aligned modes matrix shown on Fig. 4.1c left:

$$\hat{\overline{S}}_{mn} = \hat{S}_{amn} F(m,n), \tag{4.13}$$

Then, $\overline{S}(x,y)$, Fig. 4.2a, is obtained from the inverse FFT of $\hat{\overline{S}}_{mn}$. The projection of S_a on $V_\perp$ is given by $\widetilde{S} = S_a - \overline{S}$, Fig. 4.2b.



(a) **(b)**

Figure 4.2: (a) 2D-plot of the aligned part of S_a, $\overline{S}$ (b) 2D-plot of the non aligned part of S_a, $\widetilde{S}$ obtained as $\widetilde{S} = S_a - \overline{S}$

A satisfactory decomposition is obtained here, where errors related to FFT transformation ($\mathcal{L}^2 \approx 10^{-15}$, machine precision in this test) are much lower than numerical diffusion discretization.

The FFT transform allows an exact projection on $V_{b\|}$. This method is however limited in its scope: its applicability is restricted to problems with periodic solutions, uniform magnetic field direction and parallel diffusivity. In addition, the mesh must be uniform. In order to generalize the field decomposition, alternative approaches are given in the next section.

4.2.2 The field averaging method along the parallel diffusion line

In order to have access to $\overline{S}$ in bounded problems (independently of the boundary conditions), an alternative filtering method is proposed here based on an averaging along the parallel diffusion line. Indeed, the obtained average has a vanishing parallel gradient.

For any point in the domain, the averaging along the parallel diffusion direction can be obtained as follows:

$$\mathcal{F}\, S_{a,ij} = \frac{1}{N+1}\left(S_{a\,ij} + \sum_{p=1}^{N} S_p^{int}\right), \tag{4.14}$$

where N is the number of $\mathcal{X}_i$ planes intercepted by the parallel diffusion line going through the mesh point with index (i,j), S_p^{int} is the interpolated field values on the parallel diffusion line, and $\mathcal{F}\, S_{a,ij} = \bar{S}_{i\,j}$ is the calculated mean field. Note that Eq. 4.14 is independent of the grid point (i,j): it only depends on the parallel direction. Thus,

if the parallel diffusion line goes through two different grid points, Eq. 4.14 will result in the same mean on both points. Since the method provides a value for each line along the direction of $\mathbf{b}$, then $\bar{T} = \bar{S}$ is invariant to the parallel diffusion operator, leading to:

$$\boldsymbol{\nabla} \cdot (K_{b\|} \cdot \boldsymbol{\nabla} \bar{T}) = 0. \tag{4.15}$$

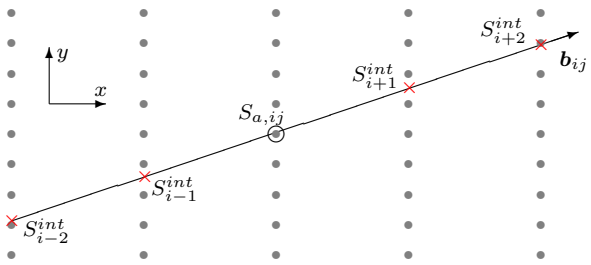


Figure 4.3: Chart of the interpolated averaging method for the parallel diffusion line defined at $S_{i\,j}$. For each intercepted y-plane p, an interpolated value S^{int}_{ap} is calculated.

The parallel diffusion direction being known in the domain, the coordinates of points where the interpolation values S^{int}_{p} must be computed can be obtained as seen in Sec. 2.3.3. The y relative coordinate y^+ in x_{i+1} reads:

$$y^+ = y + \int_{x_i}^{x_{i+1}} \frac{b_y}{b_x} dx, \tag{4.16}$$

and analogously for any integral limit $x_{i\pm p}$. The interpolated field value S^{int}_{p} is obtained here by polynomial approximations on the y-direction, Fig. 4.3. For example, in any plane $\mathcal{X}_{i+s}$, we can obtain the 3^{rd} degree centered interpolation for y^+ being between y_k and y_{k+1} as:

$$S^{int}_{i+s} = a_0 + a_1(y - y_k) + a_2(y - y_k)^2 + a_3(y - y_k)^3 \tag{4.17}$$

For any given quadruplet of grid values $(S_{a,i+s,k-1}, S_{a,i+s,k}, S_{a,i+s,k+1}, S_{a,i+s,k+2})$, Eq. 4.17 leads to:

$$\begin{bmatrix} a_0 \\ a_1 \\ a_2 \\ a_3 \end{bmatrix} = \begin{bmatrix} 1 & -\delta y & \delta y^2 & -\delta y^3 \\ 1 & 0 & 0 & 0 \\ 1 & \delta y & \delta y^2 & \delta y^3 \\ 1 & 2\delta y & 4\delta y^2 & 8\delta y^3 \end{bmatrix}^{-1} \begin{bmatrix} S_{a,i+s,k-1} \\ S_{a,i+s,k} \\ S_{a,i+s,k+1} \\ S_{a,i+s,k+2} \end{bmatrix}$$

Since this method is an explicit operation performed at all grid points, the interpolation degree does not have the drawback related to the matrix inversion and mentioned in Sec. 4.1. In the present document, tests have been carried out from linear interpolation to 6^{th} centred interpolation degree.

Considering $[F]_{N_xN_y \times N_xN_y}$ the discrete matrix of $\mathcal{F}$ and S_a in Ω_{ij}, we consider $\bar{S}$ such as:

$$[\bar{S}]_{N_x N_y} = [F]_{N_x N_y \times N_x N_y} [S_a]_{N_x N_y}, \tag{4.18}$$

Contrary to the direct method described in Sec. 4.2.1, the present method is iterative: a number of iterations, it is needed to obtain the mean, invariant field $\overline{S}$:

$$\overline{S} = \sum_{\forall it} \bar{S} \tag{4.19}$$

The algorithm to obtain the field decomposition is described in Algorithm 1:

Result: $\overline{S}$ and $\widetilde{S}$ iterative process.
initialization;
$\overline{S}^{(0)} = 0$;
for $it = 1 : it_{fin}$ **do**
  $\bar{S}^{(it)} = \bar{S}^{(it-1)} + \mathcal{F}(S_a - \bar{S}^{(it-1)})$;
end
$\widetilde{S} = S_a$ - $\bar{S}^{(it_{fin})}$;

Algorithm 1: $\overline{S}$ calculation from interpolations along the parallel diffusion direction considering it_{fin} iterations.

4.2.3 The local averaging method

As seen in the previous section, the field averaging method along the parallel diffusion line provides advantages over the filtering in modal space since it allows projection on $V_{b\|}$ in bounded domains with non-uniform field directions. The matrix F is usually sparse since values entering in the computation of the average are taken in a narrow neighbourhood around a **b** line. However, the proposed projection actually consists of a convolution along the field line, and the associated matrix is still of dimension $(N_x N_y) \times (N_x N_y)$. Its construction is thus numerically costly. Moreover, whether the matrix is constructed or not does not change the fact that the construction of the average is a nonlocal operation. A local version of the previous method is therefore proposed here.

For all discrete versions of the parallel Laplacian compared here (*Classic, Günter, Ottaviani, Stegmeir* and *Present* schemes), the stencil is restricted to 3 neighbouring planes $\mathcal{X}_{i-1}$, $\mathcal{X}_i$ and $\mathcal{X}_{i+1}$. The construction of the discrete projection operator is *local*, in the sense that the discrete operator only uses the values of the closest grid points.

Taking the same interpretation as in Sec. 4.2.2, for any grid point $S_{a\,ij}$, we can obtain a local mean to find the invariant part $\overline{S}_{ij}$ of the discrete Laplacian operator in this *local scope*. This interpretation allows us to reduce the stencil of the mean field to a size comparable to the discrete Laplacian. This opens the possibility of using a sparse matrix

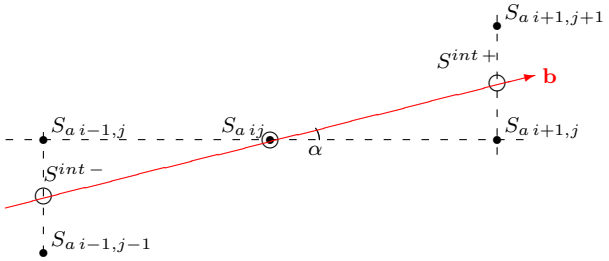


Figure 4.4: Stencil of the local averaging on $\boldsymbol{b}$ in $S_{a\,ij}$. S^{int+} and S^{int-} are values linearly interpolated along the y-direction.

representation of the filter, and the same grid size than the one used into the discrete Laplacian, Fig. 4.4.

The discrete mean of $S_{a\,ij}$ is obtained from values interpolated in the surrounding planes: $S^{int-} \in \mathcal{X}_{i-1}$ and $S^{int+} \in \mathcal{X}_{i+1}$. The two means proposed here are:

$$(\mathcal{F}_1 S_a)_{ij} = \frac{1}{3}(S^{int-} + S_{a\,ij} + S^{int+}), \tag{4.20}$$

called *local averaging method 1* (LA1), and:

$$(\mathcal{F}_2 S_a)_{ij} = \frac{1}{4}(S^{int-} + 2\,S_{a\,ij} + S^{int+}), \tag{4.21}$$

called *local averaging method 2* (LA2). Both methods are clearly idempotent for a field which is invariant along the parallel direction. They are however hampered by discretization errors, related to the interpolation used in constructing values on the neighbouring planes S^{int-} and S^{int+}.

The application of the two filters smoothes out fluctuations along field lines. The parallel average can then be defined using the filters $\mathcal{F}_1$ and $\mathcal{F}_2$ as:

$$\bar{S}_{ij} = \lim_{k\to\infty} (\mathcal{F}_1^k S_a)_{ij} \quad ; \quad \hat{S}_{ij} = \lim_{k\to\infty} (\mathcal{F}_2^k S_a)_{ij} \tag{4.22}$$

A recursive application of any of the two will converge towards the mean along the magnetic field. The fluctuations with parallel variations are then recovered using:

$$\widetilde{S}_{ij} = S_{a\,ij} - (\mathcal{F}_1^k S_a)_{ij} \quad ; \quad \widetilde{S}_{ij} = S_{a\,ij} - (\mathcal{F}_2^k S_a)_{ij} \tag{4.23}$$

4.2.4 The filtering methods based on Laplacian discretizations

The averaging methods described in the previous sections can be reinterpreted as a Laplacian solver which only resolves the parallel modes. Then, all parallel Laplacian

discretisations presented in Chapters 2 and 3 can be re-scaled to be used as a filter and becoming now a filter of the fluctuating part. Considering any of these discrete parallel Laplacian methods, here represented by the sparse matrix $\mathcal{F}$, $\widetilde{S}^{(k)}$ can be obtained as:

$$\widetilde{S}^{(k)} = \widetilde{S}^{(k-1)} + \mathcal{P}\mathcal{F}(S_a - \widetilde{S}^{(k-1)}), \tag{4.24}$$

where $\mathcal{P}$ is a re-scaling matrix to ensure that the eigenvalues of $\mathcal{P}\mathcal{F}_{Steg}$ lie in the range $]-1,1[$(here, $\mathcal{P}$ is just a diagonal matrix with $1/max(\lambda_{\mathcal{F}})$ in the diagonal). The filter now obtained retains the fluctuating part of S_a, which can be associated to an iterative process as described in Algorithm 2:

Result: $\overline{S_a}$ and $\widetilde{S_a}$ iterative process.
Initialization;
$\widetilde{S}^{(0)} = 0$;
for $it = 1 : it_{fin}$ **do**
  $\widetilde{S}^{(it)} = \widetilde{S}^{(it-1)} + \mathcal{P}\mathcal{F}(S_a - \widetilde{S}^{(it-1)})$;
end
$\bar{S} = S_a$ - $\widetilde{S}^{(it_{fin})}$;

Algorithm 2: Calculation of $\widetilde{S}$ using a re-scaled parallel Laplacian discretization $\mathcal{P}\mathcal{F}$.

4.3 Test cases

To evaluate both decomposition methods, two 2D test cases are proposed here, with an analytical source term composed by a linear combination of aligned modes along **b** (first term) and non-aligned modes (second term) such as:

$$S_a(x,y) = \sum_r C_r \cos\left(m_{y,r}y - m_{x,r}x\right) + \sum_s C_s \sin\left(m_{y,s}y - m_{x,s}x\right) \tag{4.25}$$

leading to a pitch angle $\alpha = tan^{-1}(m_{x,r}/m_{y,r})$, and with C_r and C_s being the amplitudes of the two aligned and non-aligned terms. The Fig. 4.1a shows the field defined by Eq. 4.25 for the parameters defined in Table 4.1. In Eq. 4.25, the first and the second summation term represents the invariant ($\overline{S}_a$) and the non invariant ($\widetilde{S}_a$) part along the parallel diffusion, respectively.

The domain and grid details have been already provided in Sec. 2.3.1. The considered test case solving Eqs. 4.3 and 4.2 is the same than in Sec. 3.5.3 for a 64×512 grid:

$$S_a(x,y) = \cos\left(m_y y + m_{x,1}x\right) + 0.25\sin\left(m_{x,2}x\right) \tag{4.26}$$

where $\alpha = \tan^{-1}(4/27)$, $m_y = 27$, $m_{x,1} = 4$, $m_{x,2} = 2$. In this case, the source term can be splited into the invariant, $\overline{S}_a$, and non invariant part, $\widetilde{S}_a$, as follows:

Aligned				Non-aligned			
r	C_r	$m_{y,r}$	$m_{x,r}$	s	C_s	$m_{y,2}$	$m_{x,2}$
1	1	1	3	1	1	2	2
2	2	2	6	2	1	-1	1
3	0.5	9	3	3	1	1	1

Table 4.1: Parameters used in Eq. 4.25 to obtain the test case of Fig. 4.1 with $\tan\alpha = 1/3$.

$$\overline{S}_a(x,y) = \cos\left(m_y y + m_{x,1} x\right) \tag{4.27}$$

$$\widetilde{S}_a(x,y) = 0.25\sin\left(m_{x,2} x\right) \tag{4.28}$$

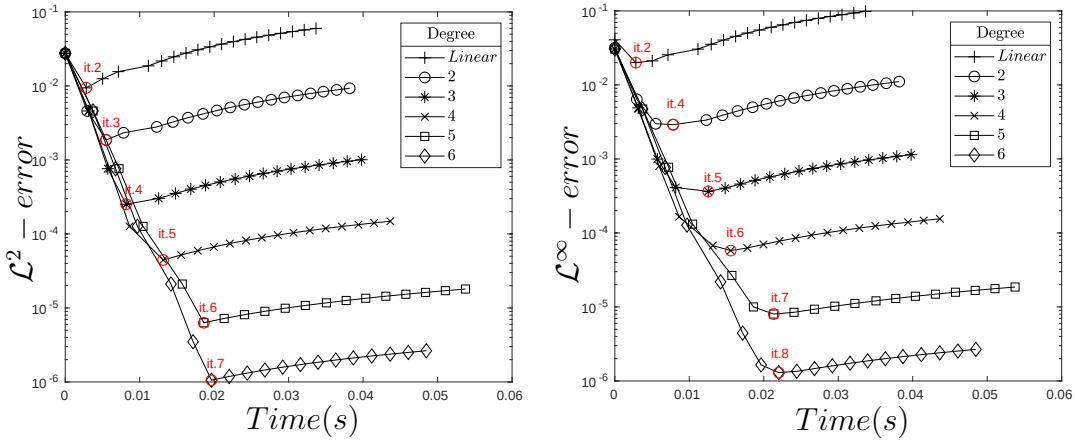
4.3.1 The field averaging method along the parallel diffusion line

Tests have been performed to evaluate the quality in estimating $\overline{S}$ for the test case of Eq. 4.25. Results for different interpolation degrees are shown with respect to both the $\mathcal{L}^2-$ and $\mathcal{L}^\infty-$ errors on Figs. 4.5a, and 4.5b, respectively. The errors obtained from both norms are comparable: the different trends represent the accumulation of the invariant part of $\overline{S}$ as $\overline{S}^{it} = \overline{S}^{it-1} + \bar{S}$ during successive iterations. The highest precision that can be achieved (red dots) depends on the degree of interpolation used in the calculations of the mean, but also on the number of iterations needed (increasing the time in the calculus). Once the highest precision is reached, the successive iterations accumulate errors related to the interpolation step only.

From a computing time point of view, the method requires the discretization of a full matrix of size $[N_x N_y \times N_x N_y]$. This constrains the number of degree of freedom to be relatively small, and hence limits that applicability to 3D simulations.

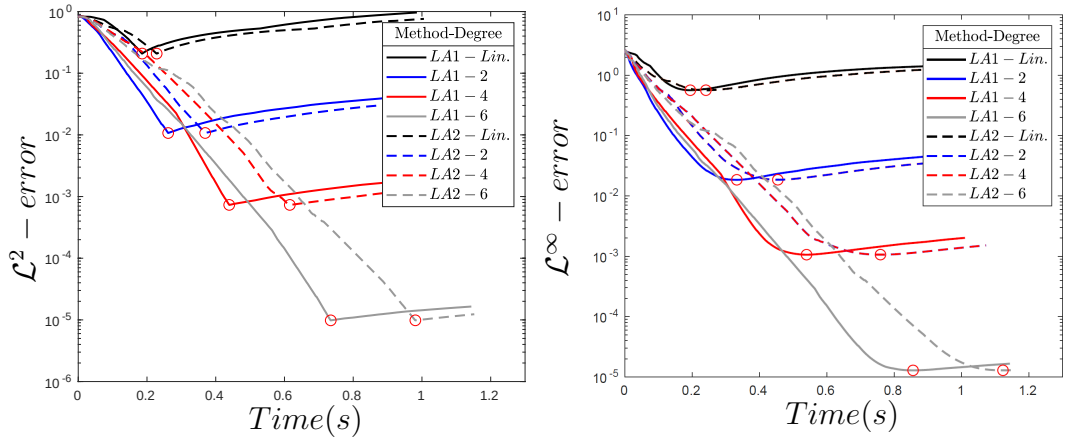
4.3.2 The local averaging method

Both $\mathcal{L}^2$ and $\mathcal{L}^\infty$ errors are shown in Fig. 4.6 for LA1 and LA2, comparing the solution $\overline{S}$ to the theoretical solution $\overline{S}_a$ given by the first term of test case 4.25. To simplify the comparison, only linear and interpolation degrees 2, 4 and 6 are considered. Results can be compared also to the field averaging method tested above, Fig. 4.5. Although LA1 and LA2 methods are faster in terms of computing time for each iteration, the global convergence towards the minimum error of the local means is slower than for the field averaging method. The global computing time is indeed at least about one order larger to obtain the same order of convergence for $\mathcal{L}^2$ and $\mathcal{L}^\infty$-errors. Compared to LA2, LA1 reaches the smallest error with a time around 25% faster than LA2 in all cases, except for the linear interpolation for which no difference in the maximal error order is found.



(a) **(b)**

Figure 4.5: Mean averaging $\mathcal{L}^2$-error (a) and $\mathcal{L}^\infty$-error (b) for successive iterations filtering $\overline{S}$. The grid is 32×256. In red, number of iterations to achieve the minimal error for each degree of the polynomial interpolation.



(a) **(b)**

Figure 4.6: Plots of the $\mathcal{L}^2$-error (a) and the $\mathcal{L}^\infty$-error (b) for successive iterations to calculate $\overline{S}$. The grid is 32×256. The red points correspond to the minimal error corresponding to the it^{th} iteration of Tables 4.2 and 4.3 for LA1 and LA2, respectively.

Comparing now the number of iterations, results in Tables 4.2 and 4.3, show that LA1 needs 25% less iterations than LA2 to reach the minimal $\mathcal{L}^2$ and $\mathcal{L}^\infty$ errors.

LA1	Linear	deg. 2	deg. 4	deg. 6
L^2	164	352	560	895
L^∞	174	461	711	1083

Table 4.2: Number of iterations needed to reach the minimal errors as function of the interpolation degree for the LA1 scheme.

LA2	Linear	deg. 2	deg. 4	deg. 6
L^2	214	470	748	1195
L^∞	233	615	949	1447

Table 4.3: Number of iterations needed to reach the minimal errors as function of the interpolation degree for the LA2 scheme.

Residual evaluation

The last filtering test aims to predict the number of iterations needed during the filtering process. The $\mathcal{L}^2$ norm tests showed in Fig. 4.6 predicts the number of iterations needed when the solution $\overline{T}$ is known. In real applications, the solution $\overline{T}$ is obviously unknown, and the number of iterations needed for the calculation of $\overline{T} = \overline{S}$ remains unknown. Actually, as seen in Fig. 4.6, an under/over-estimation of the number of iterations compromise the accuracy of this filter method.

However, all methods show a period of convergence during the initial iterations. This period of convergence can be identified by a kind of residual R evaluated at all iterations:

$$R^{it} = \frac{\|\sum_1^{it-1} \overline{S} - \bar{S}^{it}\|}{\|S_a\|}, \tag{4.29}$$

where the numerator estimates the variation during successive filtering iterations with respect to the total field in the it^{th} iteration. The use of the total field as normalization is justified here by the fact that the field cannot have an aligned term, avoiding in this case false R^{it} estimations.

In order to model a characteristic signal of an highly anisotropic flow with a moderate number of modes, the following test case is proposed:

$$S_a(x,y) = C_1 + C_2 \sum_1^{10} sin(m_x x + m_y y) + C_3 \sum_1^{20} sin(m_{x2} x + m_{y2} y) \tag{4.30}$$

with $C_1 = C_2 = C_3 = 1$, $m_x = X$, $m_y = 7X$, $m_{x2} = X_2$ $m_{y2} = 7X_3$, where X is a set of 10 random integer numbers in the range $[1, 5]$ and X_2, X_3 a set of 50 random integer numbers in the range $[-5, 5]$ with $X_2 \neq X_3$. Then, the first sum of Eq. 4.30 generates

aligned modes with a pitch angle $\alpha = \tan^{-1}(1/7)$, and the second sum an oriented but not aligned set of modes.

The evolution of the residual R^{it} factor for LA1 and LA2 is shown in Fig. 4.7 a and b, respectively. Both results show that the number of iterations needed to reach the highest accuracy in the filtering is similar, only depending on the order of interpolation. This number of iterations is determined at points where the residual stops to decrease. All methods show a 2^{nd} order of convergence given by the interpolation in the parallel direction. However, the final precision is fixed by the interpolation in the y-direction, see Table 4.4.

Let's notice that for the linear version of the LA1 and LA2 methods the residual R^{it} continues to very slightly decrease, which corresponds to a "spurious" convergence directly linked to the interpolation precision when $\overline{S}^{it+1}$ is over-evaluated.

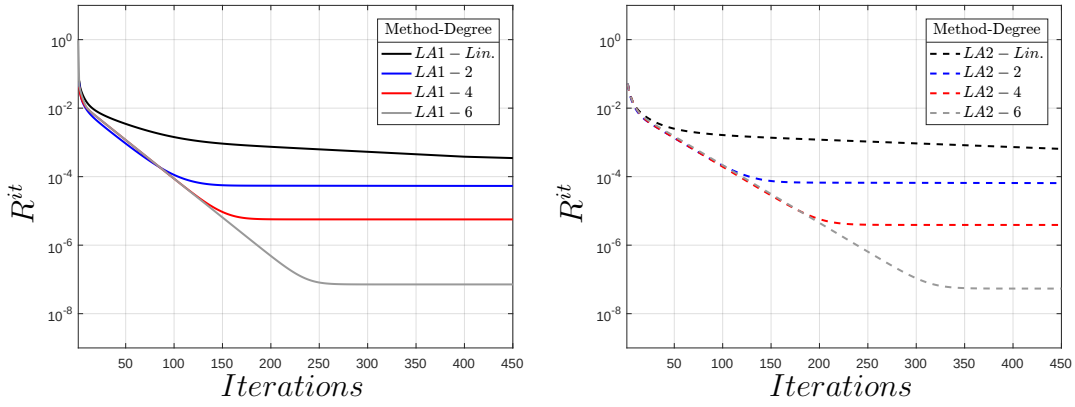


Figure 4.7: Evolution with the iterations of the residual R^{it} depending on the degree of interpolation for LA1(a) and LA2(b). The grid is 32×256.

Note that on Fig. 4.7 the initial value of the residual is around $10^{-1,7}$, which is due to the normalization term given in Eq. 4.29, $\|S_a\|$. This has no influence on the number of iterations estimated above.

	LA1-2	LA2-2	LA1-4	LA2-4	LA1-6	LA2-6
Number of iterations	155	177	201	248	277	360
Minimal residual value	$5e^{-5}$	$6e^{-5}$	$5.5e^{-6}$	$4e^{-6}$	$7e^{-8}$	$5.5e^{-8}$

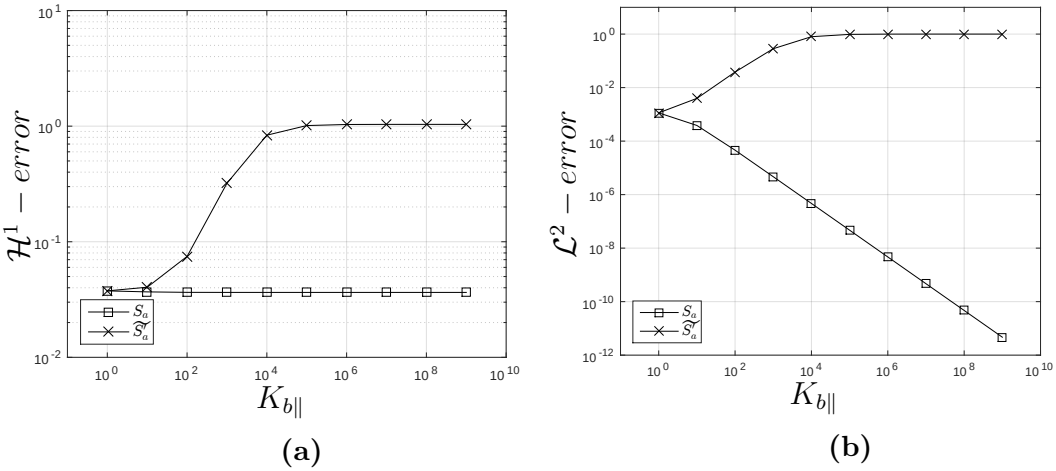
Table 4.4: Number of iterations to converge in function of the interpolation degree for the LA1 and LA2 methods, and corresponding minimal values of the residual.

4.4 Accuracy tests for the continuous problem

Tests solving the Helmholtz equation are presented here comparing the original formulation of Eq. 4.3, and the reformulated version provided 4.10. The discrete Laplacian version is given by the *Present scheme* described in Chapter 3.

In the reformulated version, the discrete operator solves only $\widetilde{S}$ after the filtering process, proving that the elimination of aligned modes has an impact on the final solution. Later on, we test if the reformulated version has also the same impact on the solution accuracy independently of the discrete Laplacian method, testing all methods presented in Chapter 2. Finally, to establish a direct comparison with results of Chapter 3, we consider the test case described in 3.5.1 in this section.

4.4.1 Filtering in modal space



(a) **(b)**

Figure 4.8: Comparisons between the $\mathcal{H}^1$-error (a) and the $\mathcal{L}^2$-error (b) for successive values of $K_{b\|}$ when solving the Helmholtz equation with the present scheme. Squared and cross points correspond to the results when solving Eq. 4.10, and Eq. 4.3, respectively. The grid is 64×512.

Results obtained with the *present scheme* are shown on Fig. 4.8. The plots of the $\mathcal{H}^1$ error shows that the precision obtained when solving the non aligned field $\widetilde{S}$ becomes independent of the value of the parallel diffusion $K_{b\|}$. The saturated value of the $\mathcal{H}^1$-error when solving the reformulated version is driven by the gradient evaluated in the $\mathcal{H}^1$ formulation, see Appendix C.

When solving the aligned modes, the source of error related to the interpolations along the aligned modes is amplified by the values of $K_{b\|}$ in the discrete operator. This leads to a spurious perpendicular diffusion already identified in tests cases shown in Chapter 3. Solving the reformulated version, gradients along the interpolation direction are reduced

compared to those given by the non-aligned modes that lead to lower gradients. Then, the values of $K_{b\|}$ do not have the amplification effect shown solving the original formulation.

The plots of the $\mathcal{L}^2$-error confirms the aligned modes as being at the origin of the numerical error. Indeed, in highly anisotropic problems, the non-aligned modes are strongly damped, and the solution is nearby completely aligned. Then the $\mathcal{L}^2$-error tends to zero when increasing $K_{b\|}$, since $\widetilde{S} \approx 0$ for the highest values of $K_{b\|}$, and therefore, the discrete Laplacian error also tends to 0.

4.4.2 The field averaging method

Based on interpolation methods, the field averaging method does not provide the same precision than the filtering in the modal space. However, the precision required to obtain $\overline{S}$ depends on the precision of the numerical Laplacian discretization for solving $\widetilde{S}$. Solving the Helmholtz equation in the *original* (Eq. 4.3) and the *reformulated* (Eq. 4.10) versions, we can establish two different sources of error related to the parallel diffusion discretization:

- When solving the original version, the source of error is related to the discrete Laplacian solving the source S_a, and leading to the error $\mathcal{E}(T)$.
- When solving the reformulated version, two sources of error can be identified: $\mathcal{E}(\overline{S}) = \mathcal{E}(\overline{T})$ related to filtering the invariant part, and $\mathcal{E}(\widetilde{T})$ related to the resolution of the invariant part by the discrete Laplacian considering the source $\widetilde{S} = S_a - \overline{S}$. Then, a satisfactory precision for the filtering is given when the global error depends on the discrete diffusion operator solving $\widetilde{T}$, i.e. when $\mathcal{E}(\overline{T}) \leq \mathcal{E}(\widetilde{T})$.

This feature is shown on tests solving the analytical solution Eq. 4.26 by the original and reformulated version using a 64×512 grid. The tests have been carried out from linear to 6^{th} interpolation degrees. The number of iterations corresponds to the smallest $\mathcal{L}^2$-error in the calculus of $\overline{S}$ shown on Fig. 4.5a.

The Fig. 4.9 shows the $\mathcal{H}^1$-error. For $K_{b\|} = 1$, the $\mathcal{H}^1$-error becomes independent of the interpolation degree, since the numerical diffusion remains small for all methods. For $K_{b\|} = 10^6$, numerical diffusion is eliminated from degree 0 to 1. The plots show that a higher interpolation degree is not needed since the $\mathcal{H}^1$-error saturates (in this case, due to the gradient estimate in the evaluation of the $\mathcal{H}^1$-error, see Appendix C).

The Fig. 4.10 shows now the $\mathcal{L}^2$-error For $K_{b\|} = 1$ the field averaging method clearly mitigate the errors of the discrete operator for all methods and they saturate around $\mathcal{L}^2 \approx 10^{-4}$. The source of error is given by the resolution of the non aligned modes by the finite-differences step (which has the same order in all the proposed methods). For higher $K_{b\|} = 10^6$, since all the non aligned modes treated by the discrete operator are damped, T becomes nearby an aligned field. Thus, in highly anisotropic cases, the precision of the solution for the resolution of the parallel modes depends on the order of

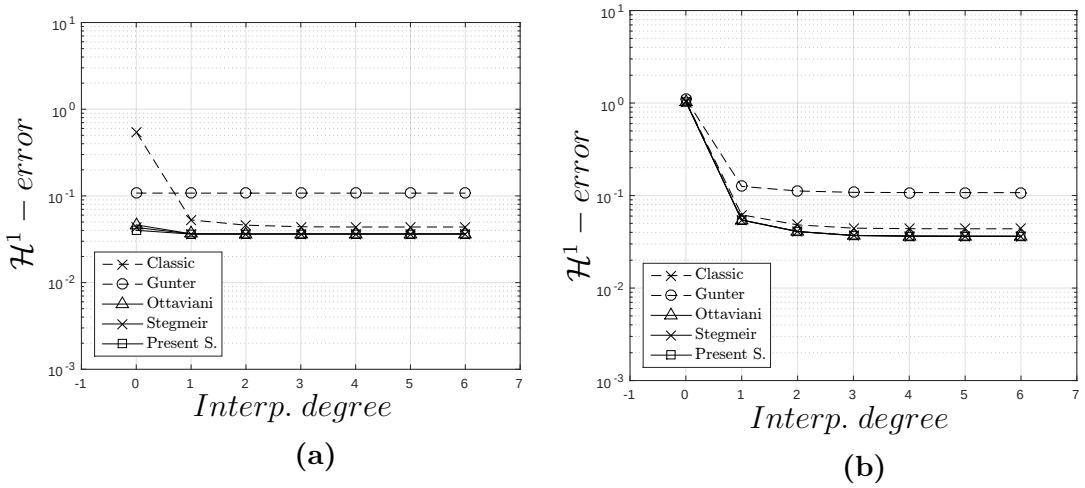


Figure 4.9: Plots of the $\mathcal{H}^1$-error with respect to the degree of interpolation used in the field averaging method to obtain $\overline{T}$ for (a) low parallel diffusion ($K_{b\|} = 1$) and (b) high parallel diffusion ($K_{b\|} = 10^6$). The interpolation degree=0 corresponds to solve Eq. 4.3 without the filtering of S_a. Bi-periodic case with the grid 64×512.

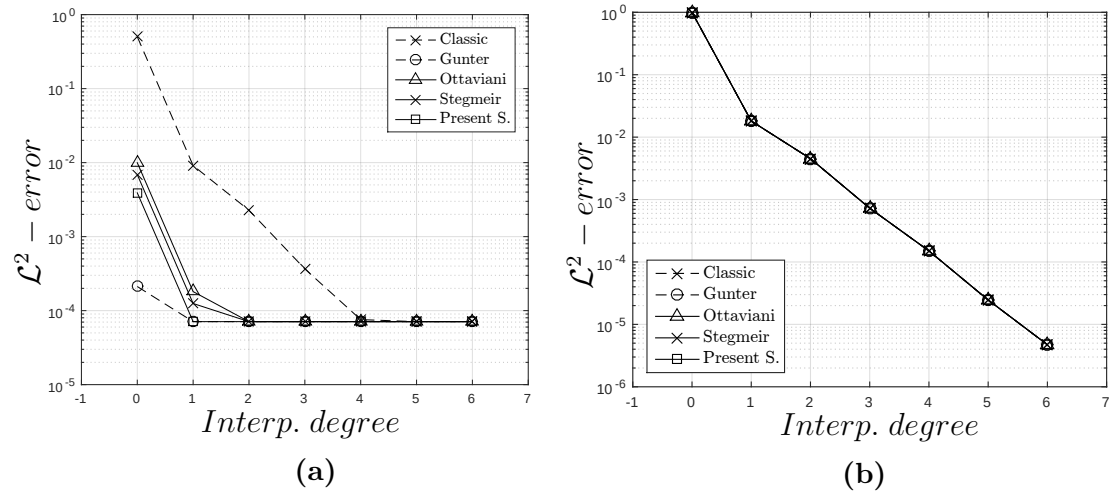
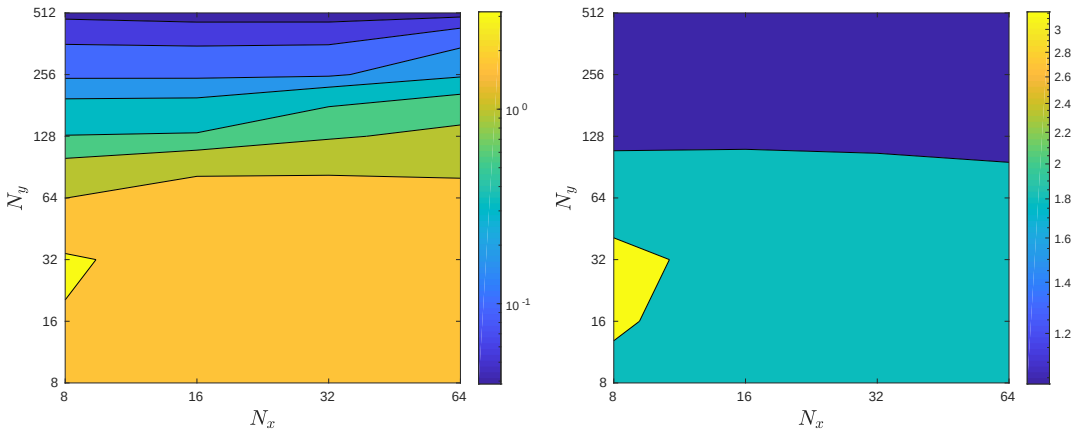


Figure 4.10: Plots of the $\mathcal{L}^2$-error with respect to the degree of interpolation used in the field averaging method to obtain $\overline{T}$ for (a) low parallel diffusion ($K_{b\|} = 1$) and (b) high parallel diffusion ($K_{b\|} = 10^6$). The interpolation degree=0 corresponds to solve Eq. 4.3 without the filtering of S_a. Bi-periodic case with the grid 64×512.

the filtering, since all non aligned modes are highly damped by the discrete Laplacian in all compared methods.

A comparison solving Eq. 4.3 (Fig. 4.11) and Eq. 4.10 (Fig. 4.12) for (a) $K_{b\|} = 1$ and (b)$K_{b\|} = 10^6$ in function of N_{dof} has been carried out for the *Present* scheme using



(a) **(b)**

Figure 4.11: 2D plots of the $\mathcal{H}^1$-error depending on N_x and N_y for (a) a low parallel diffusion ($K_{b\|} = 1$) and (b) a high parallel diffusion ($K_{b\|} = 10^6$). Bi-periodic case for the present scheme solving Eq. 4.3.

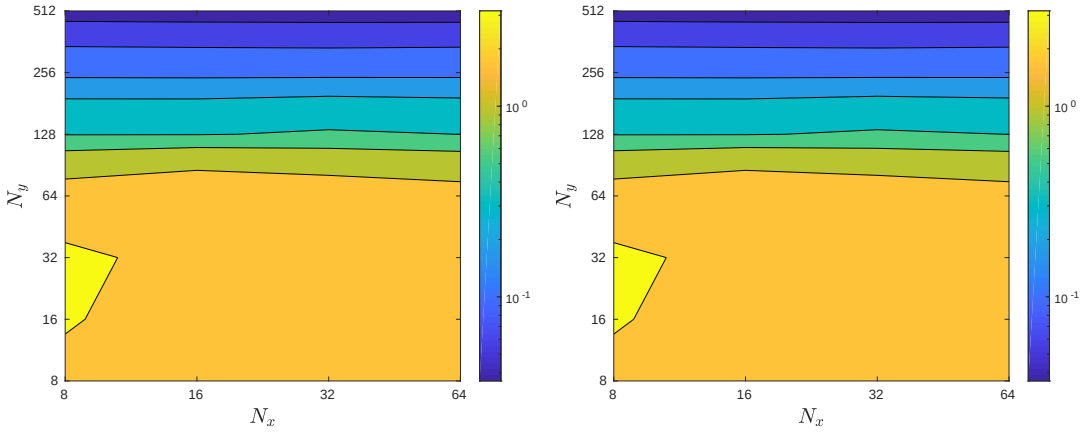
a 6^{th} degree of interpolation in Field Averaging Method for 8 iterations in the filtering process. Comparing the results for $K_{b\|} = 1$ both cases give practically the same $\mathcal{H}^1$-error tendence, since numerical pollution remains low. For $K_{b\|} = 10^6$, numerical diffusion is completely eliminated, and convergence is found with a strong reduction of N_{dof}. This fact opens to a strong N_{dof} reduction also in the y-direction for high $K_{b\|}$ values, where same order of $\mathcal{H}^1$-error are found here in about 4 times less of N_y, (see Sec. 3.5.3 for a wide range of N_{dof}).

4.4.3 The local averaging method

Due to the sparser definition of the local averaging method, the filtering application is here extended to wider ranges of N_{dof} using the LA1 method. As seen in Sec. 4.2.3, the LA1 method requires a larger number of iterations than the field averaging method to reach the accuracy of the method, this number increasing also with the number of degree of freedom N_{dof}.

Results when solving Eq. 4.3 on Fig. 4.13 are here compared with results obtained when solving the reformulated version Eq. 4.10 using the LA1 method for 1000 iterations (Fig. 4.14) and for 10000 iterations (Fig. 4.15). Results are analyzed for a low and high parallel diffusion, $K_{b\|} = 1$ and $K_{b\|} = 10^{-6}$.

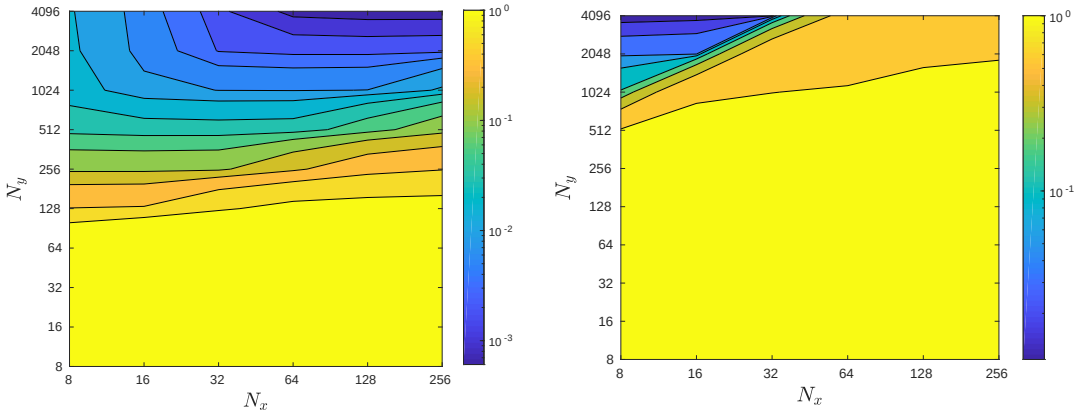
Due to the low numerical diffusion in the discrete operator for $K_{b\|} = 1$, the filtering method does not improve the solution and it is useless in this case. However, for a large value of the parallel diffusion, here $K_{b\|} = 10^6$, the filtering method eliminates the spurious diffusion as soon as the number of iterations is large enough. Indeed, when using



(a) **(b)**

Figure 4.12: 2D plots of the $\mathcal{H}^1$-error depending on N_x and N_y for (a) a low parallel diffusion ($K_{b\|} = 1$) and (b) a high parallel diffusion ($K_{b\|} = 10^6$). Bi-periodic case for the present scheme solving Eq. 4.2 with a degree 6 of interpolation for the filtering method.

the LA1 method with 1000 iterations, the estimate of $\overline{S}$ is not satisfactory, since the error rises in the region corresponding to the highest N_{dof} ($N_x > 128$ and $N_y > 256$). However, increasing the number of iterations, here to 10000, provides a satisfactory result.

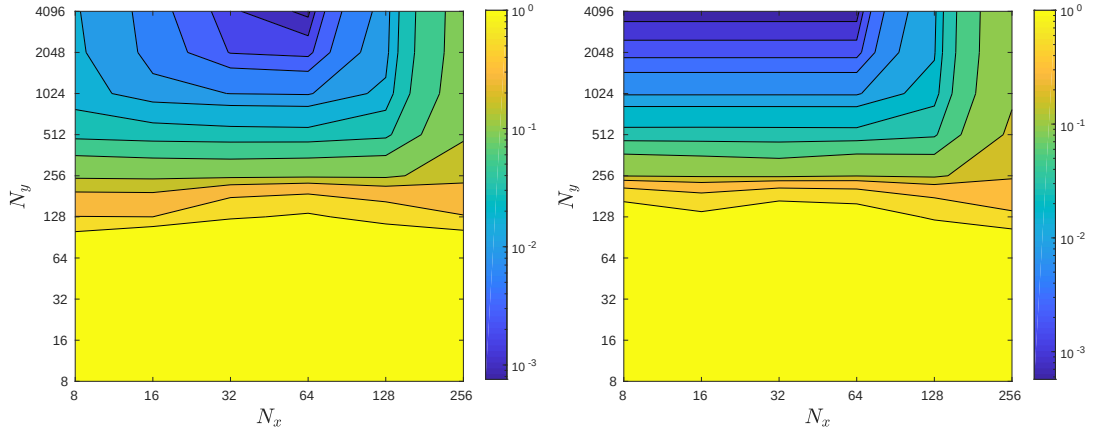


(a) **(b)**

Figure 4.13: 2D plots of the $\mathcal{H}^1$-error depending on number of grid points N_x and N_y for (a) a low parallel diffusion ($K_{b\|} = 1$) and (b) a high parallel diffusion ($K_{b\|} = 10^6$). Bi-periodic case for the present scheme solving Eq. 4.3.

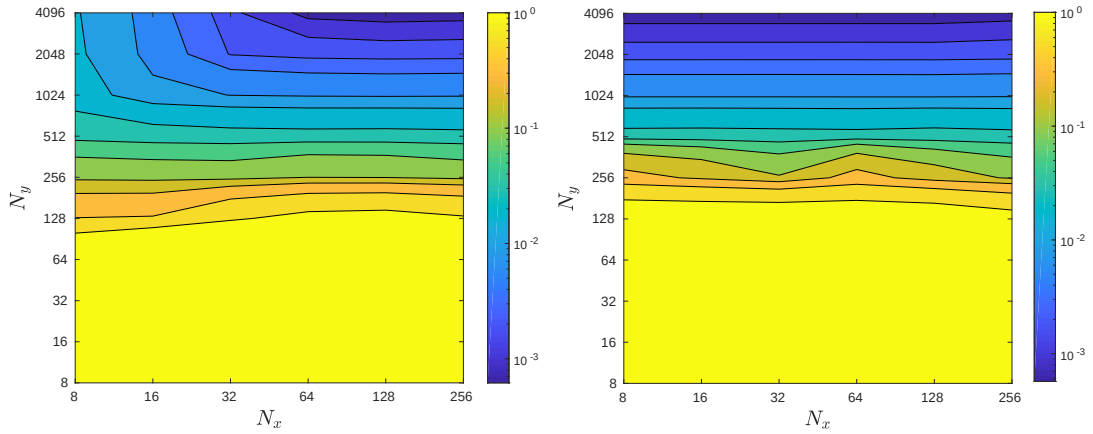
However, for $K_{b\|} = 10^6$ (Fig. 4.13b, Fig. 4.14b and Fig. 4.15b) the discrete problem converges uniformly from resolutions comparable to $K_{b\|} = 1$ in the area where filtering is applied suitably: the application of LA1 provides here a solution nearby independent

of $K_{b\|}$ value and precision depends only by N_y.



(a) **(b)**

Figure 4.14: 2D plots of the $\mathcal{H}^1$-error depending on N_x and N_y for (a) a low parallel diffusion ($K_{b\|} = 1$) and (b) a high parallel diffusion ($K_{b\|} = 10^6$).Bi-periodic case for the present scheme solving Eq. 4.2. $\overline{T}$ is obtained here by the LA1 method with 1000 iterations.



(a) **(b)**

Figure 4.15: 2D plots of the $\mathcal{H}^1$-error depending on N_x and N_y for (a) a low parallel diffusion ($K_{b\|} = 1$) and (b) a high parallel diffusion ($K_{b\|} = 10^6$).Bi-periodic case for the present scheme solving Eq. 4.2. $\overline{T}$ is obtained here by the LA1 method with 10000 iterations.

4.5 Conclusion

A decomposition of the Helmholtz equation has been presented here to limit the spurious numerical diffusion related to the discrete Laplacian for high $K_{b\|}$ values. The decomposition splits the initial field T into an invariant ($\overline{T}$) and not invariant ($\widetilde{T}$) part with respect to the parallel diffusion operator, having noticed that $\overline{T} = \overline{S}_a$ when solving the Helmholtz equation.

The modal filtering method presented in Sec. 4.2.1 allows to isolate with a high precision the aligned modes, denoted $\overline{S}_a$. However, the use of FFT limits the application of the method to bi-periodic cases.

The field averaging method proposed in Sec. 4.2.2 can be effective for in bounded problems since it is based on computing the mean along the parallel diffusion direction. By construction, the mean-field $\overline{S}_a$ is invariant along the parallel diffusion direction, the values being obtained along this direction by polynomial interpolations. The influence of the degree of interpolation has been evaluated showing that high degrees are needed to obtain an accurate estimation of $\overline{S}_a$. However, the precision needed can vary in function of the discrete Laplacian precision when solving $\widetilde{S}_a$ in the reformulated Helmholtz equation.

In order to extend the applicability of methods based on interpolations along the parallel direction to large sparse systems, the local averaging version has been presented in Sec. 4.2.3. The local averaging method has a slower convergence than the field averaging one and shows the same order of error on the $\widetilde{S}_a$ approximation. Then, an alternative to the local average methods has been proposed based on the Laplacian discretizations presented in Chapters 2 and 3. Finally, the estimate during the filtering iterations of a residual term Eq.4.29, allows determining the number of iterations needed to obtain $\overline{T}$ with the highest (or required) precision. The global efficiency of the method has been proven by solving the Helmholtz equation, Sec. 4.4, for all the proposed filtering methods. In all tests cases, the solution of the problem 4.10 reduces the spurious diffusion for any discrete Laplacian independently of the $K_{b\|}$ value. Indeed, using the aligned Laplacian methods, the solution also becomes independent of the number of points in the x-direction for high parallel diffusion, since all non-aligned modes are rapidly damped.

Chapter 5

An iterative solver for highly anisotropic elliptic problems

5.1 Introduction

In Chapters 2 and 3, five finite difference discretizations for the anisotropic Laplacian operator have been presented and compared. These discretizations have been compared in terms of precision by the method of manufactured solution (MMS), solving the following Helmholtz equation:

$$T - \boldsymbol{\nabla} \cdot (\boldsymbol{\mathcal{K}} \cdot \boldsymbol{\nabla}) T = S_a, \ \in \Omega \subset \mathbb{R}^3 \tag{5.1}$$

To generalize the scope of this chapter to any anisotropic elliptic equation, Eq. 5.1 is solved at a discrete level, where it can be written under the form of the following linear system:

$$Au = b \tag{5.2}$$

$[A]$ being a $n \times n$ matrix of discretization coefficients, u a n-vector of unknowns, and b a n-vector of source terms. For all tests presented in Chapters 2 and 3, Eq. 5.2 is solved using a Matlab inversion algorithm for unsymmetric sparse linear systems called UMFPACK, [Dav04]. Although this method is extensively used to solve system like Eq. 5.2, as it is heavily optimized for the direct solution of linear systems, the UMFPACK algorithm remains expensive in terms of memory and computing time, particularly for systems involving a large number of unknowns, which is usually the case when considering 3D multiphysics problems like the one of interest for us in a close future and presented in Chapter 1.

An alternative way to solve such a linear system Eq. 5.2 is to use iterative methods. The convergence of iterative methods is based on successive improvements of the approximated solution, modifying one or several components of the solution vector at each

iteration (*relaxation*) until the previously established criterion is satisfied. This criterion is usually based on a precision needed on the global solution.

The issue when considering a solution like S_a (Eq. 5.1) is that the iterative method may smooth certain modes more efficiently than others. As will be detailed later in Sec. 5.2, the classic iterative methods (Jacobi, Gauss-Seidel) are based on a fixed-point relaxation to solve the linear system of equations. The convergence of these iterative methods is directly related to the projected mode on a given grid, and highly oscillating modes are efficiently damped while the others are more slowly damped.

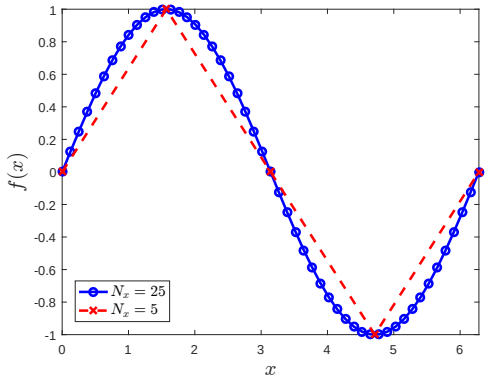
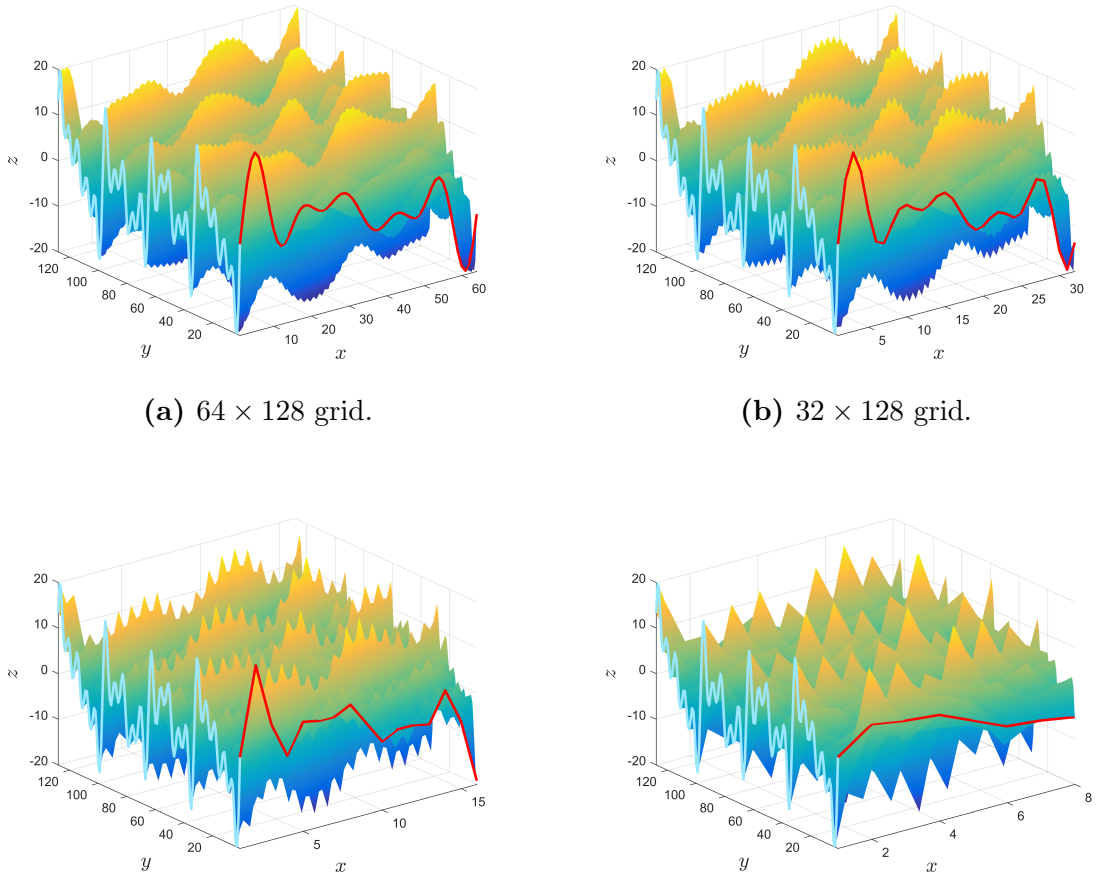


Figure 5.1: Plot of the function $f(x) = sin(2\pi x)$ projected on a grid of 25 grid (blue line) and on a grid with 5 points only (red line).

To determine if a mode has a high or a low frequency depends on the grid on which it is projected. A high-frequency mode projected on a coarse grid of N_c nodes is characterized by a high variation between successive nodes. Then, the mode exhibits a rapid variation projection in the grid. However, considering the same mode projected now on a fine grid with $N_f \gg N_c$, the variations between successive nodes are now slow, and then the mode projection can be interpreted as a low-frequency mode. This effect is shown on a example on Fig. 5.1. The function shows a slower oscillation on the fine grid than on the coarse one. This feature has an important influence when an iterative method is used to solve a linear system. Indeed, the same mode can be *relaxed* efficiently or not depending on the number of grid nodes on which the mode is projected.

In Chapter 4, projection methods have been implemented to avoid the treatment of the invariant part of the field by the discrete parallel Laplacian, eliminating thus the spurious numerical diffusion due to the treatment of the aligned modes by the discrete Laplacian. In the present chapter, a projection is proposed to optimize the relaxation step of the iterative method. A slow mode defined on a fine grid can be projected (*transfer*) into a coarse grid, becoming thus a rapid mode efficiently damped by the iterative (*smoothing*) method. The working principle of the method is then to make the smoothing rate of a given mode independent of mesh size (h-independence), in order to obtain a method whose computing time scales in optimal cases as $N_{dof} \ln N_{dof}$ [TS01].



(a) 64×128 grid.

(b) 32×128 grid.

(c) 16×128 grid.

(d) 8×128 grid.

Figure 5.2: 3D plots showing the progressive reduction of the number of grid points in the x-direction. The red line shows a view of the evolution from slow modes in fine level to rapid modes at coarse level.

In general, the source term S_a will be a composition of different modes, each mode having an optimal grid where it will be able to be rapidly damped by the smoothing method. Then, the original grid must be transferred on several coarser grids, where the smoother will damp efficiently these modes having rapid frequencies at each level. This multilevel problem, concatenating smoothing processes at each level, makes the principle of the multigrid method. The multigrid algorithm is a multilevel solver in which each mode is relaxed at the optimal level by the smoother or relaxation method. The transition between the different levels is made by transfer functions, which act as projections from the current grid to a coarser or a finer grid level.

Anisotropic elliptic problems

In a highly anisotropic diffusion case Eq 5.39, the parallel direction is characterized by slow modes in the parallel direction and rapid modes in the perpendicular one. As seen in the previous chapters, the pitch angle α is usually small in fusion applications, so the projected modes in the x-direction are significantly slower than modes projected in the y-direction. Then, a grid reduction in the x-direction can improve the efficacy of the iterative method.

The Fig. 5.1 illustrates the transition from an original fine grid (Fig. 5.1a) to coarser grids (Fig. 5.1b, c, and d) in the case of characteristic highly anisotropic field (the grid reduction is made in the x-direction). The red line shows the first line of grid nodes, showing the transition from slow to rapid modes during the reduction of the grid points number. Note that modes in the y-direction can be considered as rapid modes in the original grid and thus are supposed to be damped efficiently by the smoother method.

In the following, we present the main basis to build a multigrid method. The classic iterative methods are presented in Sec. 5.2 followed by the grid transfer methods in Sec. 5.3, where a specially adapted grid transfer is introduced here to deal with high anisotropic diffusion. Then, the multigrid algorithm is built in Sec. 5.4 comparing the different combinations of smoothing and transfer methods in terms of convergence, computing time and memory for a 3D case and a high number of *d.o.f.*. The proposed method is generalized for Dirichlet, Neumann and Robin boundary conditions in Sec. 5.4.2. Finally, in Sec. 5.5, the multigrid method introduced before is used to build a preconditioning matrix in a GMRES method. The results are compared to results obtained a with ILU(0) preconditioning for different relevant cases. For more details on the multigrid algorithms and iterative methods, the reader is referred to [BHM00, Sha95, Wes92, Saa03].

5.2 Iterative methods

The iterative methods solve the linear system 5.2 by successive iterations during which one or more components of the approximate solution vector are modified. For each iteration, *it*, the method might converge to the system solution minimizing the residual equation:

$$r^{(it)} = b - Au^{(it)} \tag{5.3}$$

Some interesting properties of the discrete matrix A defined here are favorable for the convergence of iterative methods:

- P_1: A is a sparse matrix: the number of zero elements in A is much larger than non-zero elements.
- P_2: A is symmetric (all eigenvalues are real values).

- P_3: A is positive definite : $u^T Au > 0$, $\forall u \neq 0$ (all eigenvalues are positive). Here u_n is a n-dimension vector compatible with A.
- P_4: A is diagonally dominant : $a_{ii} \geq \sum_{j \neq i} |a_{ij}|$, $\forall i$.

	P_1	P_2	P_3	P_4
Classic	×	×		
Günter	×	×	×	×
Ottaviani	×	×	×*	×
Stegmeir	×	×	×	×
Present S.	×	×	×	×

Table 5.1: Properties verified by the parallel Laplacian when using the Classic, Günter's, Ottaviani's, Stegmeir's and the present scheme (Chapter 3) for uniform diffusion terms in the parallel direction. * The Ottavani's method with a linear interpolation gives a positive definite matrix, but this property is not guaranteed as when using a SOM method.

The Table 5.2 summarizes the four properties verified or not by the parallel Lapacian discretizations depending on the methods used. It shows that all discretizations assembled by a Support Operator Method verified the four properties.

5.2.1 The Jacobi iterative method

Considering the linear system Eq. 5.2, the discrete sparse matrix A can be seen as a sum of the following sparse matrix:

$$A = L + D + U \tag{5.4}$$

where L is the lower part, D the diagonal part, and U the upper part of A. Considering $R = L + U$, Eq. 5.2 leads to:

$$Du + Ru = b \tag{5.5}$$

The solution approximated at $it + 1$ considers the solution previously found at the it^{th} iteration as:

$$u^{(it+1)} = D^{-1}(b - Ru^{(it)}) \tag{5.6}$$

leading for each element to u^{it+1}:

$$u_i^{(it+1)} = \frac{1}{a_{i\,i}} \left(b_i - \sum_{i \neq j} R_{i\,j} u_i^{(it)} \right) \tag{5.7}$$

A damped version of the Jacobi method can be obtained considering $\omega \in]0,1]$

$$u^{(it+1)} = \omega D^{-1}(b - Ru^{(it)}) \tag{5.8}$$

leading to:

$$u_i^{(it+1)} = \frac{\omega}{a_{i\,i}} \left(b_i - \sum_{i \neq j} R_{i\,j} u_i^{(it)} \right) \tag{5.9}$$

Then, Jacobi iterations are obtained considering $\omega = 1$. The effect of the damped version is studied here. According to Eq. 5.8, the iterative matrix adopts:

$$S_\omega = \mathbb{I} - \omega D^{-1} A \tag{5.10}$$

with $\mathbb{I}$ an identity matrix. The convergence of the damped Jacobi method is determined by the eigenvalues of the iteration matrix:

$$\lambda_i(S_\omega) = 1 - \frac{\omega}{2N} \lambda_i(A) = 1 - 2\omega sin^2\left(\frac{k\pi}{2N}\right) \tag{5.11}$$

Note the obtained eigenvalues are the same than the finite differences matrix A:

$$\lambda_i = \frac{4}{\Delta x} sin^2\left(\frac{k\pi}{2N}\right) \tag{5.12}$$

with $\lambda_i = \lambda_i(S_\omega) \in]-1,1[$ for all $K = 1, 2, ..., N-1$, and by considering $\{\mathbf{w}_1, \mathbf{w}_2, ..., \mathbf{w}_{N-1}\}$ assembled with by S_ω eigenvectors. The effect on the dumped modes is analyzed by studying the evolution of the error e considering:

$$\mathbf{e}^{(it)} = S_\omega^m \mathbf{e}^{(0)} \tag{5.13}$$

the error of the first iteration being:

$$e^{(0)} = \sum_{i=1}^{N-1} c_i \mathbf{w}_i \tag{5.14}$$

c_i being any real number established for the initialization. For the it^{th} iteration, the error vector $\mathbf{e}^{(m)}$ writes:

$$\mathbf{e}^{(it)} = \sum_{i=1}^{N-1} c_i S_\omega^{it} \mathbf{w}_i = \sum_{i=1}^{N-1} c_i S_\omega^{m-1}(S_\omega \mathbf{w}_i) = \sum_{i=1}^{N-1} c_i S_\omega^{it-1} \lambda_i(S_\omega \mathbf{w}_i) \tag{5.15}$$

Continuing the same decomposition for the its iterations leads to:

$$\mathbf{e}^{(it)} = \sum_{i=1}^{N-1} c_i \lambda_i^{it}(S_\omega \mathbf{w}_i) \tag{5.16}$$

Then, after it iterations, the i^{th} Fourier mode has been damped by a factor $|\lambda_i^{it}(S_\omega)|$, using the Jacobi method. On Fig. 5.3, the eigenvalues of the matrix S_ω with different damping value ω is shown. The test has been carried out considering a 1D problem solving Eq. 5.1 with the *Classic* approach. In this basic case, higher frequencies are damped efficiently for $\omega = 1/2$, the efficient range for the other values of ω being situated in other spectral locations. For example, for the classic Jacobi method with $\omega = 1$, slow and high frequencies are slowly damped, the middle frequencies being damped efficiently. However, the choice of the appropriate value of ω depends on the eigenvalues of the matrix A, which depend also from the discrete problem. Thus these results can not be extended directly to other cases.

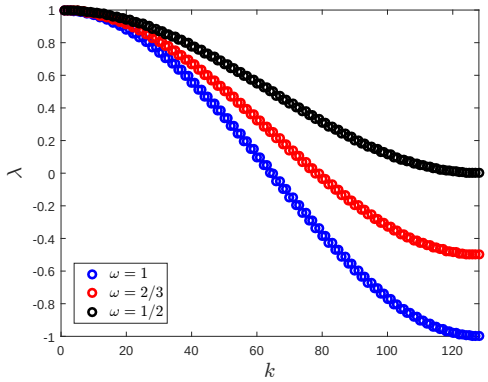


Figure 5.3: Plots of S_ω eigenvalues for a 1D case with $N_x = 128$ when solving Eq.5.1 for three values of ω.

5.2.2 The Gauss-Seidel iterative method

As described for the Jacobi method, A can be seen as a sum of several matrix, Eq. 5.4. Then, considering $M = L + D$ from 5.2 we write:

$$\begin{aligned} (M+U)u &= b \\ u &= M^{-1} + (b - Uu). \end{aligned} \tag{5.17}$$

We can obtain a recursive process from Eq. 5.17, which converges to the solution of the system Eq. 5.2:

$$u^{(it+1)} = M^{-1}b - M^{-1}Uu^{(it)} = u^{(it)} + M^{-1}(b - Au^{(it)}) \tag{5.18}$$

where u^{i+1} and and u^i are the solutions obtained at the iteration $i+1$ and i, respectively. This makes a recursive process of approximations of the system solution. Each element of vector u writes:

$$u_i^{(it+1)} = u_i^{(it)} + \frac{1}{a_{i\,i}} \left(b_i - \sum_{j=1}^{i-1} a_{i\,j} u_j^{(it+1)} - \sum_{j=i}^{n} a_{i\,j} u_j^{(it)} \right) \tag{5.19}$$

Convergence

The error given by the it^{th} iteration is denoted by:

$$\varepsilon^{(it)} = u - u^{(it)} \tag{5.20}$$

Considering $\|S^{(it)}\|$ the contraction number for the it^{th} iteration ($\|S^{(it)}\|$ is the norm of matrix $S^{(it)}$, which depends on the iterative method and the discrete operator matrix A), we establish from Eq. 5.20:

$$\|\varepsilon^{(it)}\| \leq \|S^{(it)}\| \cdot \|\varepsilon^{(0)}\| \tag{5.21}$$

Then the iterations converge when $\lim_{it\to\infty}\|S^{(it)}\| = 0$. This results can be interpreted from a spectral point of view: Eq. 5.21 converges if and only if $\mu(S) < 1$, being $\mu(S) = max|\lambda_i(S)|$ the maximal S eigenvalue (the spectral radius of S, $\rho(S)$). For the Jacobi iterative method, S is obtained by rewriting Eq. 5.6 as:

$$u^{(it+1)} = u^{(it)} + D^{-1}(b - Au^{(it)}) \tag{5.22}$$

with $b = Au$, and u is the exact solution of the linear system. The error at each iteration reads $\varepsilon^{(it)} = u^{(it)} - u$. Then:

$$\varepsilon^{(it+1)} = (\mathbb{I} - D^{-1}A)\varepsilon^{(it)} \tag{5.23}$$

$\mathbb{I}$ being the identity matrix. We establish $S^{(it)}_{Jac} = (\mathbb{I} - D^{-1}A)$, for which the spectral radius must be < 1. An equivalent convergence condition is established by the inequality $\rho(S) \leq \|S\|$: if the matrix norm of S is ≤ 1, the method converges (which is the case for all the proposed discretizations).

5.2.3 Damped modes by the iterative methods

The general idea about using a multigrid routine to solve the linear system Eq. 5.2 is illustrated in the next test case. Given the wavenumber k, we propose the 2D r.h.s defined in $]0, 2\pi[$ in x and y as:

$$b(x, y) = \sin(k\pi x) \tag{5.24}$$

The wavenumbers characterize the waves in a given grid: high and low wavenumbers determine highly and slowly oscillating modes, respectively. For a given k, the discrete representation of the r.h.s. test case has the following Fourier discrete analog in a $N \times N$ grid:

$$b(n, m) = \sin\left(\frac{k\pi n}{N}\right) \tag{5.25}$$

being $n = 1, ..., N-1$. The factor k/N in Eq. 5.25 determines if the wavenumber has a slow or rapid oscillation nature in the grid: having k a rapid discrete analog in a $N_1 \times N_1$ grid, the same k have a slow discrete analog in a $N_0 \times N_0$, being $N_0 \gg N_1$. This fact has a strong influence on the convergence of the iterative methods. To show this, the r.h.s. Eq. 5.24 with $k = 4$ (Fig. 5.4a) is tested solving the bi-periodic Helmoltz equation by the Jacobi and the GS iterative methods for the same conditions as described in Sec. 3.5.1 (Present S. Laplacian discretization, $\alpha = 0$, $K_{b\parallel} = 1$, $K_{b\perp} = 0$).

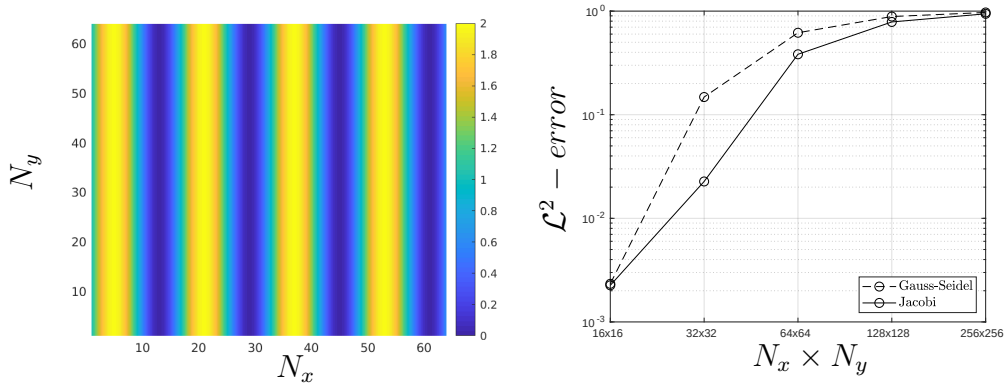


Figure 5.4: 2D plot of the r.h.s test case for $k = 4$ in Eq. 5.24 (a). Plot of the $\mathcal{L}_2$-error for the Jacobi (continuous line) and the Gauss-Seidel (dashed line) mode smoothing, for the r.h.s. plotted in (a), after 100 iterations, and for different grid sizes when solving the Helmholtz equation with the the present method.

The Fig.5.4b shows a rapid convergence for both iterative methods in reduced grids, where the wavenumber is perceived as a high-frequency wave and rapidly damped. In dense grids, the iterative methods converge slowly to the solution. In a general framework, the solution is a sum of all the Fourier modes defined on the grid:

$$b(x_n, y_m) = \sum_{k_1=1}^{N-1} \sum_{k_2=1}^{M-1} \sin\left(\frac{2\pi k_1 n}{N} + \frac{2\pi k_2 m}{M}\right), \tag{5.26}$$

where $k_1 = 1, ..., N-1$ and $k_2 = 1, ..., M-1$ are the modes defined in each grid direction. For any given $(N \times M)$ grid, lower $1 \leq k_1/N \leq k_{1,max}/(2N)$ and $1 \leq k_2/M \leq k_{2,max}/(2M)$ relation represent the modes with a slower convergence. If we consider the same wavenumbers but now in a $(N/2 \times M/2)$ grid, the upper part of the slower modes of $N \times M$ becomes rapid modes in $(N/2 \times M/2)$ grid, where iterative methods converge rapidly.

5.3 The grid transfer methods

In Sec. 5.2, we concluded there is a $N_i \times N_i$ grid where a given mode can be efficiently treated by the iterative methods. The present section describes how to transfer the information between different grids in order to solve the linear system as a multilevel (or multigrid) problem.

This transfer of information is made by interpolations from the coarser to the finer grid. The interpolation matrix is called *prolongation*, in the sense it allows to pass from $\Omega^{2\Delta x}$ defined in a $N/2 \times M$ grid, to $\Omega^{\Delta x}$ domain defined in a $N \times M$ grid, in the case of an interpolation in the x-direction.

In highly anisotropic flows, with the parallel diffusion direction not aligned to the cartesian grid, and with a small pitch angle, the solution is characterized by rapid modes in the perpendicular direction, and by slow modes in the parallel direction ($k_y \gg k_x$). Due to this feature, the perpendicular modes (rapid modes) can be considered as efficiently treated by the iterative methods. Thus, two projection methods for a resolution reduction are proposed here to reduce the grid in the x-direction (where the slower modes are found).

5.3.1 Interpolation in the x-direction

The first prolongation method proposed here is classically used in multigrid algorithms for finite difference methods. Considering the previous spaces $\Omega^{2\Delta x}$ (defined in N/2 intervals) and $\Omega^{\Delta x}$ (defined in N intervals), we can define the interpolation matrix $P^{\Delta x}_{2\Delta x}$ to obtain $\Omega^{\Delta x}$ from $\Omega^{2\Delta x}$ elements as:

$$\begin{aligned} u^{\Delta x}_{2i\,j} &= u^{2\Delta x}_{ij} \qquad i = 1, ..., N-1, \\ u^{\Delta x}_{2i+1\,j} &= \frac{u^{2\Delta x}_{ij} + u^{2\Delta x}_{i+1j}}{2} \qquad i = 2, ..., N, \end{aligned} \tag{5.27}$$

The resulting $P^{\Delta x}_{2\Delta x}$ matrix is a $N/2M \times NM$ sparse *prolongation* matrix:

$$P_{2\Delta x}^{\Delta x} = \frac{1}{2}\begin{bmatrix} 1 & 0 & \cdots & \cdots & 0 \\ 2 & & & & \vdots \\ 1 & 1 & & & \vdots \\ \vdots & 2 & & & \\ & 1 & \ddots & & \\ & & \ddots & 1 & \\ & & \ddots & 2 & \vdots \\ \vdots & & & 1 & 1 \\ \vdots & & & & 2 \\ 0 & \cdots & \cdots & \cdots & 1 \end{bmatrix} \tag{5.28}$$

Then, the interpolated grid is obtained as:

$$u_{2i\,j}^{\Delta x} = P_{2\Delta x}^{\Delta x} u_{ij}^{2\Delta x}, \tag{5.29}$$

being $u_{ij}^{2\Delta x}$ a $N/2 \times M$ matrix, and $u_{2i\,j}^{\Delta x}$ the obtained $N \times M$ matrix in fine level.

The present interpolation method along the x-direction gives a suitable transfer between the levels. In highly anisotropic flows not aligned to the Cartesian grids, the parallel and perpendicular modes projected along the x-direction generate modes of slightly higher frequency than the ones found in the parallel direction. Then, the transfer operation between levels generates high frequencies due to the numerical error coming from the prolongation step. However, all these spurious frequencies are expected to be rapidly damped by the smoothing method. To test the influence of this feature, another prolongation method is proposed in the next section, in order to compare the influence of the reconstruction direction.

5.3.2 Parallel interpolation

The modal analysis made in Sec. B.2 in Appendix B shows that the influence of the high wavenumbers in low resolutions yields to poor quality in an eventual interpolation: higher resolutions are needed to obtain a satisfactory interpolation for high wavenumbers (we assume here the accomplishment of the Nyquist-Shannon theorem).

Since the posed prolongation step only increases the grid in the x-direction, a parallel interpolation approach matrix is proposed here, aligned with the parallel diffusion direction, Fig. 5.5. Since the slower modes are found to be parallel to the main diffusion direction, an adapted transfer method of the problem can increase the quality of the interpolation, that might result in a better global convergence of the multigrid algorithm.

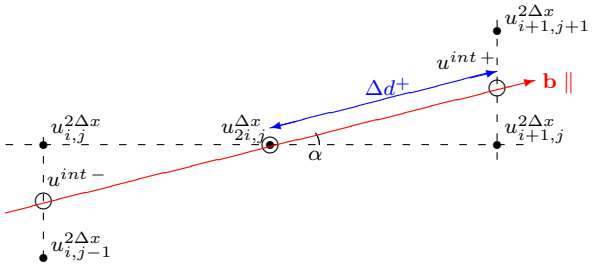


Figure 5.5: Chart of the aligned interpolation made to obtain the fine grid value $u^{\Delta x}_{2i,j}$ from the nearest points of the coarse grid to the diffusion parallel direction. u^{int+} and u^{int-} are the values interpolated in the y-direction.

The interpolation stencil is assembled from two successive 1D interpolations: given the parallel direction passing on the fine grid interpolated point, $u^{\Delta x}_{2i,j}$, it intercepts the coarser grid $\mathcal{X}^{2\Delta x}_i$ and $\mathcal{X}^{2\Delta x}_{i+1}$ planes in:

$$y^+ = y + \int_{x^{\Delta x}_{2i}}^{x^{2\Delta x}_{i+1}} \frac{b_y}{b_x} dx, \quad y^- = y + \int_{x^{\Delta x}_{2i}}^{x^{2\Delta x}_{i}} \frac{b_y}{b_x} dx, \tag{5.30}$$

Then, a linear interpolation in the y-direction is built at the y^+ and y^- positions. Considering $\xi = 0$ for simplicity (see Sec. 5.4.2), the linear interpolation leads to:

$$u^{int+} = \left(1 - \frac{y^+}{\Delta y}\right) u^{2\Delta x}_{i+1,j} + \frac{y^+}{\Delta y} u^{2\Delta x}_{i+1,j+1} \tag{5.31}$$

The expression is analogous for u^{int-}. The second interpolation is made along the parallel direction that leads to the interpolated value $u^{\Delta x}_{2i,j}$ in the fine grid:

$$u^{\Delta x}_{2i,j} = \frac{1}{\Delta d^+ + \Delta d^-}(\Delta d^+ u^{int-} + \Delta d^- u^{int+}) \tag{5.32}$$

with:

$$\Delta d^+ = \left| \int_{x^{\Delta x}_{2i}}^{x^{2\Delta x}_{i+1}} \frac{\sqrt{b_x^2 + b_y^2 + b_z^2}}{|b_x|} \, \mathrm{d}x \right|, \tag{5.33}$$

The details to obtain $\Delta d^\pm$ and $y^\pm$ have been described in Sec. 2 for the aligned methods.

The obtained stencil leads to a $N/2M \times NM$ sparse matrix. Since aligned transfer is defined by a 5-points stencil, the obtained discrete matrix is denser than the interpolation matrix in the x-direction described in 5.3.1 (defined by a 3-points stencil). This feature can lead to slower transfer steps for the aligned method, whose influence will be tested later. In terms of memory, the matrix product in the prolongation operation Eq. 5.29 with the aligned method does not seem to be problematic.

5.3.3 The reduction matrix

The prolongation matrix presented above transfers the signal from a coarser grid to a finer grid. In a multigrid routine, the transfer is made in both directions: a *reduction* matrix is required to transfer the signal from a fine grid to a coarse grid.

This reduction can be done directly by an *injection* matrix: a coarse grid is obtained directly from points selected from the fine grid. Then, the injection matrix is a sparse matrix with ones as non-zero elements:

$$R_{\Delta x}^{2\Delta x} = \begin{bmatrix} 1 & 0 & 0 & \cdots & \cdots & \cdots & 0 \\ 0 & 0 & 1 & & & & \vdots \\ \vdots & & & \ddots & & & \vdots \\ \vdots & & & & 1 & 0 & 0 \\ 0 & \cdots & \cdots & \cdots & 0 & 0 & 1 \end{bmatrix} \tag{5.34}$$

Then, the reduced grid matrix is given by:

$$u_{i\,j}^{2\Delta x} = R_{\Delta x}^{2\Delta x} u_{2ij}^{\Delta x}, \tag{5.35}$$

with $u_{i\,j}^{2\Delta x} \in u_{2ij}^{\Delta x}$.

The injection matrix is effective to reduce the grid, but the obtained coarse grid can be seen as an elimination of d.o.f. of the fine grid, so the injection is not considered as a transfer matrix, in the sense that information from eliminated points transforms the original system to another one.

The alternative to the injection matrix is the weighted reduction matrix, which takes the information from all the points from the fine grid to assemble the coarse grid by averaging:

$$u_{ij}^{2\Delta x} = \frac{u_{2i-1j}^{\Delta x} + 2u_{2ij}^{\Delta x} + u_{2i+1j}^{\Delta x}}{4}, \tag{5.36}$$

The $NM \times N/2M$ sparse matrix obtained considering this average reads:

$$R_{\Delta x}^{2\Delta x} = \frac{1}{4} \begin{bmatrix} 1 & 2 & 1 & 0 & \cdots & \cdots & \cdots & \cdots & 0 \\ 0 & 0 & 1 & 2 & 1 & & & & \vdots \\ \vdots & & & \ddots & \ddots & \ddots & & & \vdots \\ \vdots & & & & 1 & 2 & 1 & 0 & 0 \\ 0 & \cdots & \cdots & \cdots & \cdots & 0 & 1 & 2 & 1 \end{bmatrix}, \tag{5.37}$$

It has an important connection to the x-direction prolongation method (Eq. 5.28):

$$R_{\Delta x}^{2\Delta x} = 2\left(P_{2\Delta x}^{\Delta x}\right)^T, \tag{5.38}$$

This relation is extended to the aligned interpolation approach, Eq.5.32, taking the same stencil to average the coarse grid values.

5.3.4 The prolongation-reduction test

In terms of precision, the two transfer methods presented in this section are compared to a simply prolongation-reduction routine, for a given initial fine grid. The test reduces the initial resolution of 64×512 with two levels in order to obtain a 16×512 grid, and then it prolongs twice to the initial resolution. The field used for the test is described in Sec. 3.5.1, Fig. 5.6a.

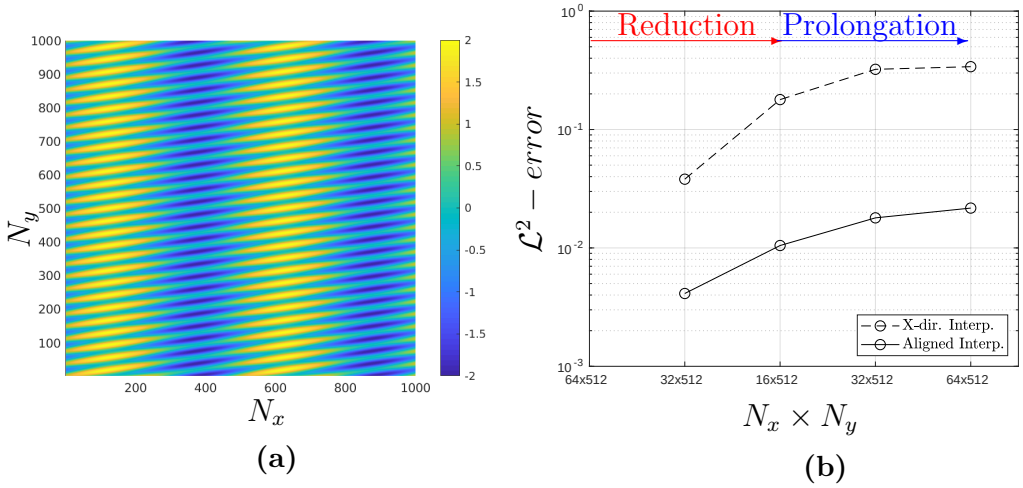


Figure 5.6: Prolongation reduction test considering an anisotropic field. (a) 2D plot of the source term in the N_x N_y plane. Plot of the $\mathcal{L}^2$-error for the two reductions ($64 \times 512 \rightarrow 32 \times 512 \rightarrow 16 \times 512$) and the two prolongations for the interpolation in the x-direction and the aligned interpolation.

Results on Fig. 5.6b, show that the aligned prolongation-reduction leads to a $\mathcal{L}^2$-error one order smaller than the values obtained for the interpolation in the x-direction, using the same degree of interpolation. These results are due to the better treatment provided by the aligned method to take into account the orientation of the highly anisotropic field, with a smaller wavenumber along the parallel direction than the one found in the x-direction.

However, these results do not assure a better performance in terms of convergence when it is used in the multigrid routine. In both methods, the error is concentrated in the reduced and interpolated points between two levels. The spurious modes generated by the grid transfer introduces high frequencies in both methods, ($k \approx N_x$ in each reduction-prolongation process). These high frequencies are expected to be rapidly damped by the smoothing method, according to the results obtained in Fig. 5.4.

5.4 The multigrid algorithm

In Sec. 5.2, we concluded that iterative methods only solve the highest frequencies efficiently. In Sec. 5.3, a transfer method has been proposed to reduce a fine grid ($N \times M$) to a coarse grid ($N/2 \times M$, or 1^{st} level), for which the slower modes can be damped more efficiently. Nevertheless, these lower wavenumbers can remain slow modes at a lower level. Then, through successive grid reductions ($N/4 \times M$, 2^{nd} level), slower modes can be treated efficiently by the smoothers. Combining this process with successive reduction-prolongation transfers leads to the multigrid structure chart shown on Fig. 5.7.

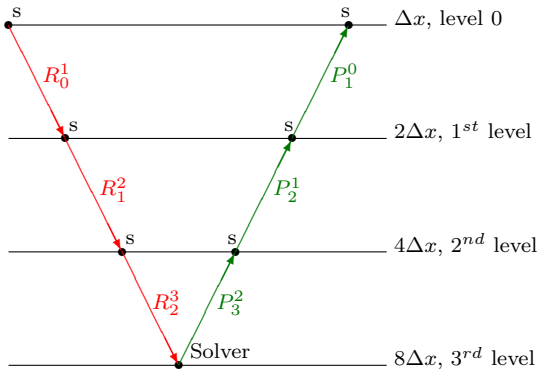


Figure 5.7: Sketch showing a 3 level multigrid V-cycle: s-smoothing, R-restriction, P-prolongation and an exact solver in the higher (coarser) level.

The process described can be seen as a cycle. A cycle is a set of reduction $[R]_n^{n+1}$ and prolongation $[P]_{n+1}^n$ steps, combined with a smoothing processes at each level. In the lower level (coarser grid), since the number of *d.o.f.* is low, a direct solver is usually applied to solve the system, which provides a better convergence of the global algorithm. The shape of the cycle is also an object of study. The cycle proposed in Fig. 5.7 is known as V-cycle. Others shapes can be found in the literature (W-cycle, F-cycle) aiming to optimize the algorithm in terms of convergence and memory. In the present work, only this V-cycle is considered in the multigrid algorithm. The tests of the other options are left to future works.

The Algorithm 3 shows the multigrid process, reducing the initial grid at the upper level, and smoothing at each level. Note that the initial grid resolution, and the reduction rate between levels, gives the maximal number of levels, can be reduced the system to one point (lowest accessible level).

5.4.1 The periodic test cases

In the previous sections, the different tools used to build a multigrid routine have been described. However, for a given problem, the ideal multigrid routine convergence depends

Data: Transfer matrix [R], [P], discrete Laplacian [A] and r.h.s vector [b]
Result: Solution vector [u] solving the linear system $Au = b$ by a multigrid V-cycle
Initialization ;
for *cycles it=1,2,...,end* **do**
  for *level l=1,2,...,p* **do**
    $A^{(l+1)} = R^{(l)}\, A^{(l)} (R^{(l)})^T \leftarrow$First restriction discrete matrix;
    $b^{(l+1)} = R^{(l)}\, b^{(l)} \leftarrow 1^{st}$ Restriction r.h.s.;
    $u^{(l+1)} = S(u^{(l+1)}) \leftarrow$ Smoothing;
  end
  $u^{(p)} = (A^{(p)})^{-1} b^{(p)} \leftarrow$ Exact solver at coarsest level;
  for *l=p,p -1,...,1* **do**
    $u^{(l)} = P^{(l+1)} u^{(l+1)} \leftarrow$ Prolongation;
    $u^{(l)} = S(u^{(l)}) \leftarrow$ Smoothing
  end
end

Algorithm 3: Algorithm of the multigrid V-cycle with p-levels to solve the linear system Eq. 5.2. The discrete operator in the final level is inverted by the UMFPACK algorithm.

on the optimal combination between them, that is not so evident. We compare here the multigrid performance in terms of computational time and memory with the aim to find the best combination for highly anisotropic diffusion, in function of:

- the final level of the V-cycle,
- the prolongation-reduction methods,
- the smoothing methods, and
- the discrete operator scheme.

In the following tests, a multigrid V-cycle is used considering complete cycles (V-cycle loops, Algorithm 3) with one smoothing step after any prolongation step.

The domain of application is a 3D grid for solving the Helmholtz equation with $\mu = 1$:

$$\begin{cases} -\boldsymbol{\nabla} \cdot \boldsymbol{\mathcal{K}} \boldsymbol{\nabla} T + \mu T = S & \text{in } \Omega, \\ \beta \nabla_{b\|} T + \gamma T = g & \text{on } \Gamma, \end{cases} \tag{5.39}$$

with $\beta = \gamma = 0$ in periodic domain. The Helmholtz equation (with $\mu > 0$) gives an invertible discretization for all proposed Laplacian schemes in Sec. 5.2 in periodic boundary conditions. The initial discrete domain (level 0 in 5.7) is defined by a $N_x \times$

$N_y \times N_z = 64 \times 512 \times 16$ grid, similar to typical resolution used in the TOKAM3X code ($64 \times 512 \times 32$). The multigrid resolution at each level is obtained by reducing the resolution in the $x-$direction, with $N_x = 32, 16, 8, 4, 2, 1$ through the successive lowest levels.

In the following tests, we consider the parallel diffusion in the $x - y$ plane like in TOKAM3X (where all magnetic surfaces, i.e. magnetic field lines, are contained in a poloidal-toroidal plane), the perpendicular diffusion being here assumed to be zero, $K_{b\perp} = 0$. The z-direction (the radial direction in TOKAM3X) is assumed to be aligned with the grid. Then, the discrete finite difference Laplacian in the z-direction reads:

$$\nabla \cdot (K_z \cdot \nabla T_{ijk}) \approx \frac{K_{z\,ijk}}{\Delta z^2} \left(T_{ijk-1} - 2T_{ijk} + T_{ijk+1}\right), \tag{5.40}$$

with $K_z = 1$ in all the following tests. Then, the non aligned schemes solve:

$$-\boldsymbol{\nabla} \cdot \left[\boldsymbol{\mathcal{R}} \begin{bmatrix} K_{b\|} & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & K_z \end{bmatrix} \boldsymbol{\mathcal{R}}^{-1} \right] \begin{Bmatrix} \partial T/\partial x \\ \partial T/\partial y \\ \partial T/\partial z \end{Bmatrix} + \mu T = S. \tag{5.41}$$

$\boldsymbol{\mathcal{R}}$ being the rotation matrix of Eq. 5.41 defined as:

$$\boldsymbol{\mathcal{R}} = \begin{bmatrix} \cos\alpha & -\sin\alpha & 0 \\ \sin\alpha & \cos\alpha & 0 \\ 0 & 0 & 1 \end{bmatrix}, \tag{5.42}$$

On the other hand, the aligned schemes solve:

$$-\boldsymbol{\nabla} \cdot \begin{bmatrix} K_{b\|} & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & K_z \end{bmatrix} \begin{Bmatrix} \partial T/\partial b_{\|} \\ \partial T/\partial b_{\perp} \\ \partial T/\partial b_z \end{Bmatrix} + \mu T = S. \tag{5.43}$$

The r.h.s. term S_a corresponds to one defined for the test case described in Sec. 3.5.1 in the z-direction, (obtaining a constant field in the z-direction):

$$S_a(x, y, z) = C_1 + C_2 \cos\left(m_y y + m_{x,1} x\right) + C_3 \sin\left(m_{x,2} x\right). \tag{5.44}$$

All tests have been performed with $\alpha = \tan^{-1}(4/27)$, $m_y = 27$, $m_{x,1} = 4$, $m_{x,2} = 2$, $C_1 = 3$, $C_2 = 1$, $C_3 = 0.25$. The results are analyzed by showing the convergence of the residual elimination at each iteration when solving the linear system:

$$Residual^{it} = \sqrt{\sum_{\forall i,j} (b_{i,j} - A u_{i,j}^{it})^2} \tag{5.45}$$

Test of the V-cycle final level.

The first test determines the V-cycle optimal level in the residual elimination, Eq. 5.45. Figs. 5.8 and 5.9 compare different sizes of V-cycle (see Fig. 5.7) for the lowest V-levels (Level 4 for $N_x = 8$ as lowest level, L.5 for $N_x = 4$, L.6 for $N_x = 2$ and L.7 for $N_x = 1$). The test has been carried out for the Günter's scheme (not aligned) and the present scheme (aligned), considering the aligned transfer described in Sec. 5.3.2 and the Gauss-Seidel smoothing.

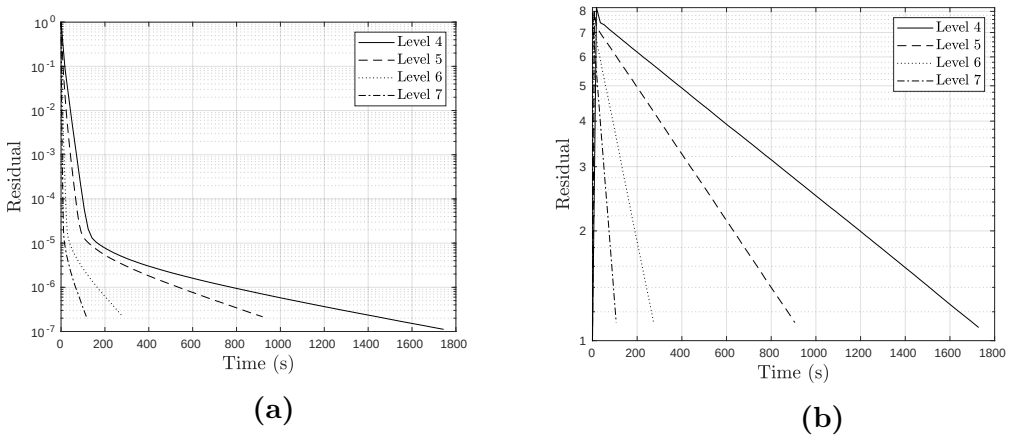


Figure 5.8: Plot of the time evolution of the residual showing its damping for the Günter's method. (a) $K_{b\|} = 1$ and (b) $K_{b\|} = 10^6$. The final grid reduction at the highest level is reduced from $N_x = 64$ the original resolution to 8 (level 4), 4 (level 5), 2 (level 6) and 1 (level 7).

With regard to the V-cycle final level, both discrete diffusion operators (Figs. 5.8 and 5.9) presents a faster convergence for the Level 7, where x-direction grid points are reduced to 1, independently of $K_{b\|}$ value.

Nevertheless, the present scheme (Fig. 5.9) achieves a faster residual reduction compared to the Günter's scheme (Fig. 5.8), leading to a total residual elimination for $K_{b\|} = 1$ and a satisfactory residual elimination for $K_{b\|} = 10^6$ before the saturation. This slower residual elimination when using the Günter's scheme can not be a consequence of the Laplacian discretization since the transfer method is favourable to aligned discretizations. In this case, the interpolation in the x-direction can be favourable to the non-aligned scheme.

Test of the transfer scheme

We test here the influence of the two transfer schemes in the multigrid V-cycle: the interpolation method in the $x-$direction (Sec. 5.3.1) and the aligned interpolation method (Sec. 5.3.2). Since in the test of the prolongation-reduction, Fig. 5.6, the aligned interpolation clearly works better, the use of a 5-point stencil regarding the 3-point stencil of the

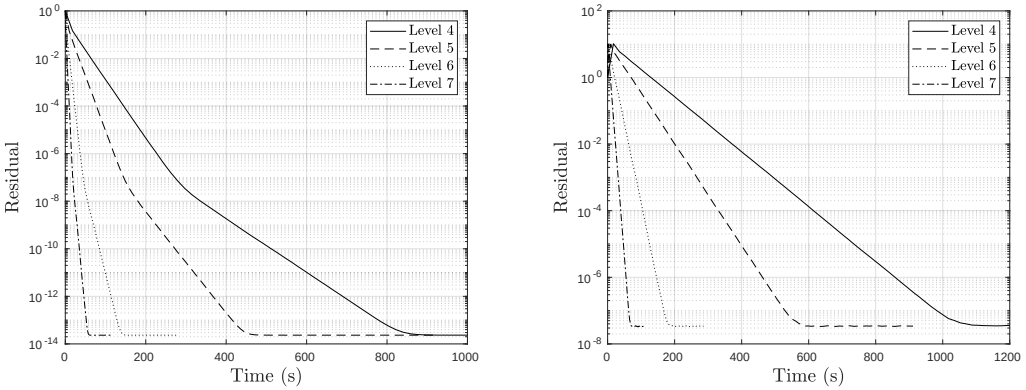


Figure 5.9: Plot of the time evolution of the residual showing its damping for the present scheme. (a) $K_{b\|} = 1$ and (b) $K_{b\|} = 10^6$. The final grid reduction at the highest level is reduced from $N_x = 64$ the original resolution to 8 (level 4), 4 (level 5), 2 (level 6) and 1 (level 7).

x-direction interpolation leads to a denser matrix in lowest levels. A direct consequence is that a larger memory is needed, that also leads to larger inversion times on the lowest levels, Table 5.2. On the other hand, the use of non-aligned transfer methods may have a positive influence on non aligned schemes.

Matrix filling rate (%)

	x-direction interpolation			Aligned interpolation		
Discretization	Level 5	Level 6	Level 7	Level 5	Level 6	Level 7
Classic	0.045	0.061	0.061	0.284	0.677	1.578
Günter	0.045	0.061	0.061	0.284	0.677	1.578
Ottaviani	0.064	0.085	0.085	0.247	0.604	1.482
Stegmeir	0.064	0.085	0.085	0.266	0.641	1.501
Present S.	0.082	0.110	0.110	0.284	0.677	1.575

Table 5.2: Fraction of non zero elements (%) for the interpolation in the x-direction and for the aligned interpolation, for levels 5 ($N_{x,final} = 4$), 6 ($N_{x,final} = 2$) and 7 ($N_{x,final} = 1$) and for an initial grid $64 \times 512 \times 16$.

The Fig. 5.10 shows the influence of previous transfer methods for a 7-level multigrid using the Günter's and the present schemes. Results show a better residual elimination per cycle for the aligned transfer method since the residual elimination for the transfer method with the interpolation in the x-direction is shown to be very slow. Even in terms of time, the aligned transfer method convergence slope is higher than the x-direction interpolation transfer. This results contrast with the sparsity of the result matrix A,

reduced from different levels: a less sparse matrix does not have any influence on the global convergence, Table 5.2. Indeed, the good performance of the aligned transfer method opens the way to an effective smoothing by the GS algorithm to improve the global convergence.

For the aligned transfer, the Günter's and the present schemes show two different slopes of convergence, since different modes are damped at different rates by each discrete method. The combination of an aligned Laplacian scheme with an aligned prolongation is clearly superior for this test case.

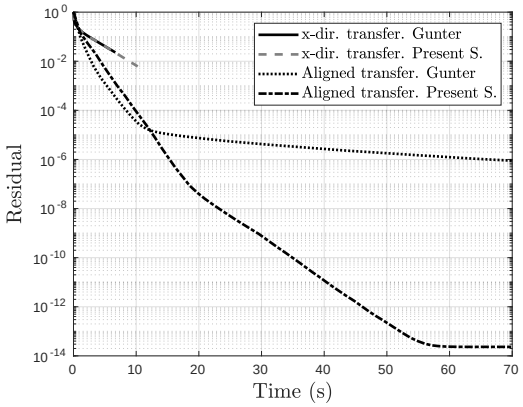


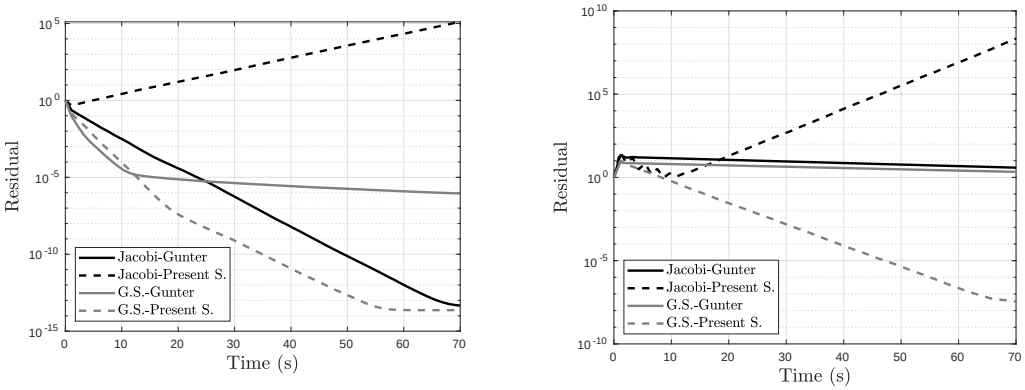
Figure 5.10: Time evolution of the residual after 100 V-cycle loops for the x-direction transfer and for the aligned interpolation transfer using the Günter's and the present schemes.

Smoothing test

Both classic smoothing methods are compared here: the Jacobi (Sec. 5.2.1) and the Gauss-Seidel (Sec. 5.2.2) methods. The test has been carried out for the 7^{th} level of a V-cycle, considering the aligned transfer method for the Günter's and the present schemes for $K_{b\|} = 1$, Fig. 5.11a, and $K_{b\|} = 10^6$, Fig. 5.11b. Only one smoothing loop after each prolongation or reduction step is consider here.

For $K_{b\|} = 1$, results show a better convergence of the Jacobi method when using the Günter's scheme, but it is not effective when using the present scheme. The Gauss-Seidel smoothing is shown to converge for both cases, the highest frequencies being damped rapidly, whereas the lowest frequencies are damped slowly (presenting different slopes at t$\approx$ 10 for the Günter's and at t$\approx$ 20 for the present scheme). Indeed, with the Günter's scheme, the convergence of the GS is less effective than with the Jacobi.

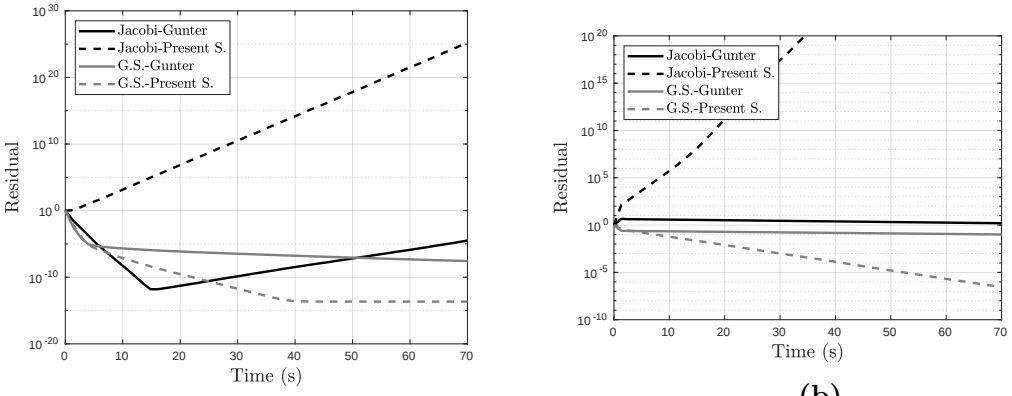
For $K_{b\|} = 10^6$, results show that only the GS smoothing converges with the present scheme, increasing the time in the residual elimination by 20%. For the Günter's scheme, the two smoothers show a very slow convergence.



(a) **(b)**

Figure 5.11: Time evolution of the residual for the Günter's and the present schemes using the Jacobi (black lines) and the Gauss-Seidel (grey lines) smoothing methods. Results are shown for level 7 in the V-cycle. The smoothing step iterates 5 times after each prolongation or reduction. (a) $K_{b\|} = 1$ and (b) $K_{b\|} = 10^6$.

To test the influence of the smoothing after each transfer step, the previous test is done now considering 5 smoothing iterations after a prolongation or a reduction step. Results of Fig. 5.12 show that several smoothing iterations do not improve the decrease of the residual for the Gauss-Seidel algorithm, independently of the $K_{b\|}$ value. Only the combination of the Günter's scheme with the Jacobi smoother improves the results for $K_{b\|} = 1$, Fig. 5.12a. However, the oversized V-loops for the Jacobi method leads to spurious accumulation in the residual, independently of the value of $K_{b\|}$.



(a) **(b)**

Figure 5.12: Time evolution of the residual using the Günter's and the present schemes and the Jacobi (black lines) and Gauss-Seidel (grey lines) smoothing methods. (a) $K_{b\|} = 1$. (b) $K_{b\|} = 10^6$. Results for the level 7 in the V-cycle.

Computing time evaluation

The computing time tests have been carried out to compare the multigrid with UMFPACK (Matlab direct solver). After the results obtained for the previous test cases, the multigrid algorithm is built with the aligned transfer, the Gauss-Seidel smoothing (1 iteration), and the present scheme as Laplacian discretization. The timing is determined by the Matlab routine for both cases (*tic-toc* function). Results shown in Table 5.3 show a better performance of UMFPACK for coarser grids, the finest grid being unsolvable due to the high memory requirements of the solver algorithm (higher than 15 Gb). Since the multigrid method used here does not improve the timing for coarse grids, the curve trend expects a time reduction for the finer grids and for low and high parallel diffusions, Fig. 5.13a and b, respectively.

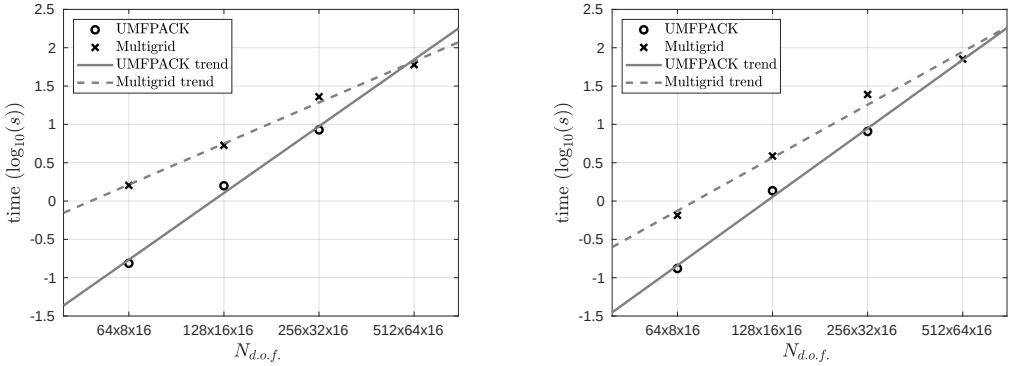


Figure 5.13: Times for solving the linear system Eq. 5.2 for the present scheme using the Matlab solver (UMFPACK) and the v-cycle multigrid for the minimal residual term. The Matlab solver for the $512 \times 64 \times 16$ grid is out of memory ($>$ 15Gb).

	UMFPACK		Multigrid	
N_{dof}	$K_{b\|\|} = 1$		$K_{b\|\|} = 10^6$	
8,192	0.154	0.132	0.601	0.652
32,768	1.582	1.363	5.335	3.861
161,072	8.483	8.092	22.932	24.533
524,288	-	-	60.151	71.041

Table 5.3: Computing time requirements for the UMFPACK and the multigrid solvers for $K_{b\|\|} = 1$ and $K_{b\|\|} = 10^6$. Results are presented for different grids sizes. UMFPACK for 524288 $N_{d.o.f.}$ is out of memory ($>$ 15Gb)

Memory tests

The benefits of the multigrid routine are evident in terms of the memory used during the process. In Table 5.4.1, the memory usage is given for different final levels and for the different Laplacian schemes. There are no substantial differences between them.

	x-direction interpolation			Aligned interpolation		
Discretization	Level 5	Level 6	Level 7	Level 5	Level 6	Level 7
Classic	3.43	3.98	4.29	5.80	5.28	6.55
Günter	3.44	3.98	4.31	6.52	5.57	6.88
Ottaviani	3.76	4.24	4.68	4.93	5.22	6.02
Stegmeir	4.10	4.25	4.71	5.55	5.80	6.67
Present S.	5.87	5.22	4.68	6.01	5.70	6.92

Table 5.4: Memory consumption [GB] for a V-cycle multigrid solver and for different final levels, transfer methods and discrete Laplacian methods and for $N_{d.o.f} = 524288$.

Nevertheless, the two transfer methods presented in Sec. 5.3 present an important difference in memory consumption up to $> 50\%$ in some cases. In general terms, the multigrid algorithm compared with the Matlab-UMFPACK (which demands > 15Gb) demands around a 70% less of memory usage, solving a $N_{d.o.f} = 524288$ system.

5.4.2 Tests in bounded domain

Bounded domains are very relevant in many engineering applications. In particular, for fusion, the simulations in the edge plasma require to deal with solid wall boundaries as introduced in Chapter 1. It is so very relevant in this thesis to extend here also the multigrid algorithm to deal with bounded computational domain. The same 3D Helmholtz problem introduced in the previous Section (Eq. 5.39) is considered here with Dirichlet, Neumann or Robin conditions in the y-direction, and keeping the periodic conditions in the x and z directions. The solid wall in the y-direction models the limiter, the main solid plasma facing component intercepted by the magnetic field line in the plasma edge. Results are related to the resolution of the linear system in a V-cycle multigrid routine.

According to the former conclusions drawn from the periodic case, we already know that:

- In periodic domain, the use of the aligned transfer combined with the aligned approach, is shown to be superior to the non aligned Laplacian and the x-direction transfer scheme. Then, for bounded domain tests, we will only consider the aligned Laplacian with the aligned transfer schemes.

- As seen in Sec. 3.4, the aligned Laplacian discretizations require an alternative discretization to lead with the limits of the domain, since stencils oriented with respect to the parallel direction can reach regions located outside the domain.
- The same limitation is presented for the aligned transfer in the bounded domain: an alternative aligned interpolation is needed near the boundary points, since the stencil is also oriented along the parallel direction, see Fig. 5.15a.

According to all these remarks, three aligned transfer methods are proposed here, depending on the treatment of the boundary. To deal with the boundary, an aligned Laplacian method is presented dealing with points near the boundary and specially adapted to the proposed transfer methods.

5.4.3 Aligned transfer methods in bounded domain

To establish the different methods in the bounded domain, we start introducing the aligned transfer matrix described in Sec. 5.3.2 as a part of the global transfer matrix. All grid points considered far from the limits will be interpolated by the aligned transfer method (points whose aligned transfer stencil is projected inside the domain). Establishing the prolongation matrix $P_{2\Delta x}^{\Delta x}$, it is equal to the interpolation matrix in the periodic conditions:

$$P_{2\Delta x}^{\Delta x} = \left[I_{2\Delta x}^{\Delta x} \right] \tag{5.46}$$

$P_{2\Delta x}^{\Delta x}$ being not a squared matrix. Considering now a bounded domain, the prolongation matrix also contains the treatment of the boundary points. To distinguish them from the unknowns at the inner domain, the treatment of the boundary points is located in a specific sector (matrix $[B]$) of $P_{2\Delta x}^{\Delta x}$:

$$P_{2\Delta x}^{\Delta x} = \left[\begin{array}{c:c} I_{2\Delta x}^{\Delta x} & C \\ \hdashline BC & B \end{array} \right] \tag{5.47}$$

Now, the $[N_x N_y, N_x N_y]$ sparse matrix $P_{2\Delta x}^{\Delta x}$ is defined by composing 4 matrices: the $[N_x(N_y-2), N_x(N_y-2)]$ interpolation matrix $[I_{2\Delta x}^{\Delta x}]$ (which is used for the transfer between inner points and redefined here to deal with bounded domains), the $[2N_x, 2N_x]$ matrix $[B]$ (which refers to the boundary points), the $[N_x(N_y-2), 2N_x]$ matrix $[BC]$ (which connects the boundary to the inner points trough the boundary conditions) and the $[2N_x, N_x(N_y-2)]$ matrix $[C]$ (which connects the inner points to the boundary conditions). In this section, three different methods are described depending on the treatment of the boundary points that leads to several definitions of $[B]$, $[BC]$ and $[C]$.

Method 1

The most simplified method is the elimination of the boundary points: boundary conditions are defined in the discrete Laplacian at level 0, and are eliminated in all the following multigrid levels, Fig. 5.14. Then, the definition of the matrix 5.47 simplifies since $[B]$, $[BC]$ and $[C]$ are the zero matrices:

$$P_{2\Delta x}^{\Delta x} = \begin{bmatrix} I_{2\Delta x}^{\Delta x} & 0 \\ 0 & 0 \end{bmatrix} \tag{5.48}$$

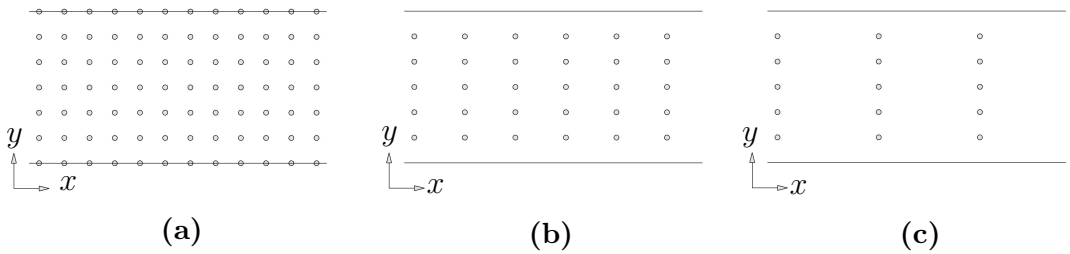


Figure 5.14: Chart of the successive restriction processes made by *Method 1* to eliminate the boundaries from level 0 (a), level 1 (b), and level 2(c). The inverse process, from (c) to (a), corresponds to the prolongation step.

As seen for the bounded Laplacian discretization, Sec. 5.4.2, the use of an aligned approach with oriented stencil requires an alternative scheme during the approximation of the boundaries. Then, the aligned interpolation scheme used to define $[I_{2\Delta x}^{\Delta x}]$ is modified for points whose stencils intercept outside the boundaries. For straight parallel diffusion lines in Cartesian structured grids, the term *shift*, previously introduced, defines the number of d.o.f lines which are overlapped by the oriented stencil projection; and then, the number of lines near the boundaries whose stencils must be modified, Fig. 5.15a.

The alternative stencil proposed as *Method 1* is based on extrapolations from inner points: given the coordinates of a point to extrapolate (prolongation step) the diffusion line defined at this grid points are projected to the inner grid points, intercepting several $\mathcal{X}$ planes, Fig. 5.15b. Then, T_{ij} being an interpolated grid point near the lower boundary (in $y = 0$), and the parallel diffusion direction $\mathbf{b}$ being straight and parallel, the extrapolation from neighbouring points writes:

$$T_{ij} = \frac{3}{2} T_{int}^{+} - \frac{1}{2} T_{int}^{++} \tag{5.49}$$

T_{int}^{+} and T_{int}^{++} being two values interpolated from inner grid points in the $\mathcal{X}_{i+1}$ and $\mathcal{X}_{i+2}$ planes, respectively. Plane overlapping and interpolation in this case is analogous to Eqs. 5.32 and 5.33.

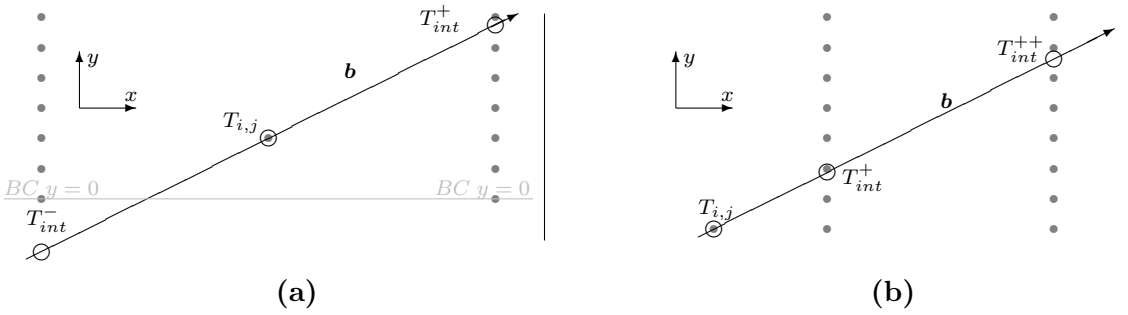


Figure 5.15: Interpolation near the boundary in $y = 0$ using the presented aligned interpolation. (a) *Aligned method* with ghost points $T^{int,-}$ located along the parallel direction and possibly far outside the domain (extrapolated points method). (b) Present method with ghost points T^{add} added on the boundary of the domain.

Remark The prolongation matrix 5.48 is established at level 0 in the V-cycle multigrid routine. For the following levels (once the boundary points are eliminated), the prolongation matrix $P^{\Delta x}_{2\Delta x}$ adopts the values and size of $[I^{\Delta x}_{2\Delta x}]$ with the extrapolations near the boundary, Eq. 5.49.

Method 2

The second transfer method maintains the original boundary points of level 0 in all the following levels: during all the reduction steps, the border points remain constant in all levels, Fig. 5.16. This method allows the discrete Laplacian to access to the boundary conditions in all levels, which can accelerate the global convergence of the method.

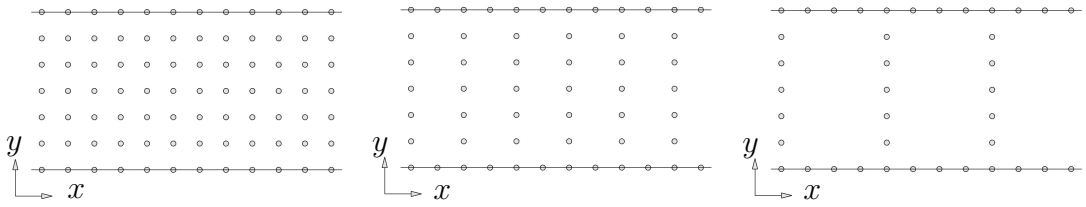


Figure 5.16: Chart of successive restriction processes made by *Method 2* and *3* maintaining the original resolution at the boundary limits from level 0 (a), level 1 (b), and level 2(c). The inverse process, from (c) to (a), corresponds to the prolongation step.

Comparing with the previously described *Method 1*, the structure of the $P^{\Delta x}_{2\Delta x}$ matrix conserves the same $[I^{\Delta x}_{2\Delta x}]$ matrix, $[B]$ being now the identity matrix $\mathbb{I}$ with a constant size for all levels equal to ($2N_x \times 2N_x$, with N_x the number of unknowns in the x-direction and at level 0):

$$P_{2\Delta x}^{\Delta x} = \left[\begin{array}{c|c} I_{2\Delta x}^{\Delta x} & 0 \\ \hline 0 & \mathbb{I} \end{array}\right] \tag{5.50}$$

Method 3

The third considered method also considers the boundary points in the interpolation process: the points extrapolated in *Methods 1* and *2* are now interpolated from inner points in one side, and from the boundaries in the other side (depending on the lower ($y = 0$) or upper ($y = 2\pi$) limit). Near the lower limit, Fig. 5.17, the value T_{ij} is interpolated as:

$$T_{ij} = \frac{1}{d^- + d^+}\left(d^- T_{int}^+ + d^+ T_{int}^-\right), \tag{5.51}$$

$'+'$ being the values calculated from inner points, Eqs. 5.32 and 5.33, and the $'-'$ being the values of the projection of $\mathbf{b}$ on the boundaries. Here, T_{int}^- is interpolated in the x-direction from neighboring boundary points in:

$$x = i\Delta x + \int_0^{y_i} \frac{b_x}{b_y} dy \quad \text{at} \quad y = 0, \qquad \text{and} \qquad x = i\Delta x + \int_{y_i}^{2\pi} \frac{b_x}{b_y} dy \quad \text{at} \quad y = 2\pi. \tag{5.52}$$

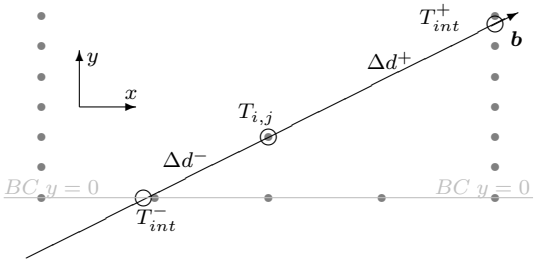


Figure 5.17: Interpolation *Method 3* near the boundary: T_{int}^- is interpolated in the x-direction from the neighboring ghost points, T_{ij} being lately interpolated from T_{int}^- and the inner interpolated value T_{int}^+.

Finally, the connection between the inner points and the boundary points is made by non-zero terms in matrix $[C]$, being the prolongation matrix $P_{2\Delta x}^{\Delta x}$ of *Method 3*:

$$P_{2\Delta x}^{\Delta x} = \left[\begin{array}{c|c} I_{2\Delta x}^{\Delta x} & C \\ \hline 0 & \mathbb{I} \end{array}\right] \tag{5.53}$$

Alternative Laplacian discretizations in bounded domains

The tests presented here are related to the resolution of the Helmholtz equation for Dirichlet boundary conditions. Numerical details are those given in Sec. 5.4.1. The grid remains structured at the boundary and the initial resolution in the x-direction satisfies the Nyquist-Shannon condition to avoid aliasing effects ($N_x = 64$ for 4 modes in the test case 5.44). Thus the Laplacian discretization near the boundary presented in Sec. 5.4.2 (which adds a grid point T^{add} at the boundary aligned to the parallel direction of T_{ij}) is modified here: the point is obtained by an interpolation in the x-direction in the x-coordinate given by Eqs. 5.52 for both boundaries.

Then, the complete discrete operator considers the same finite-difference stencil than the one presented in Sec. 5.4.2 for the added points method. Then, the flux between the inner grid point T_{ij}, $(q^-)_{y=0}$, and the corresponding interpolated grid point at the boundary $y = 0$, is obtained as:

$$(q^-)_{y=0} = \frac{T_{ij} - (T^-_{int})_{y=0}}{d^-}, \tag{5.54}$$

$(T^-_{int})_{y=0}$ being obtained by interpolation. Note that the $'+'$ formulation is similar for the boundary at $y = 2\pi$. For simplicity, considering straight diffusion lines with an orientation angle α, see Fig. 5.18b, the interpolation coordinate at $y = 0$ solving Eq. 5.52 is:

$$(x^-_{int})_{y=0} = \frac{j\Delta y}{\tan\alpha}, \tag{5.55}$$

Then, the expression of linear interpolation from boundary points leads to:

$$(T^-_{int})_{y=0} = \frac{1}{\Delta x}\left((\Delta x - x^-_{int})T_{i-1,1} + x^-_{int}T_{i,1}\right). \tag{5.56}$$

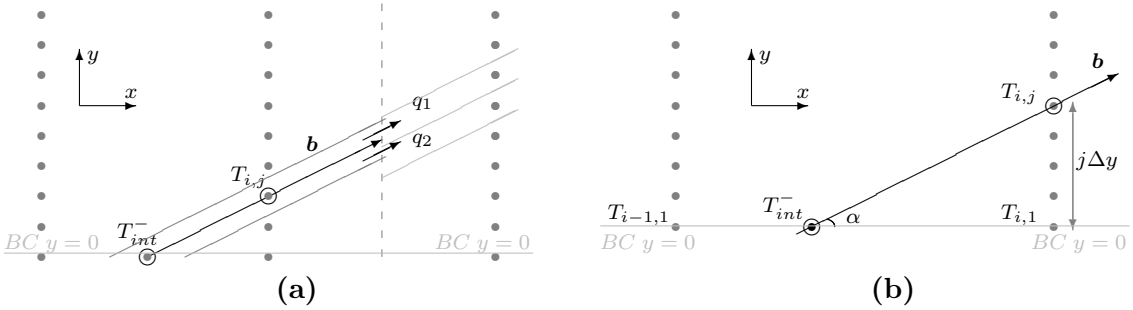


Figure 5.18: Finite-difference stencils near the boundary. (a) T^-_{int} is interpolated in the x-direction from neighboring ghost points. (b) the geometrical position of T^-_{int} is obtained by the interception of $\mathbf{b}$ defined in T_{ij} with the boundary limit.

Once the flux $(q^-)_{y=0}$ or $(q^+)_{y=2\pi}$ is obtained, the rest of the discrete Laplacian operator is similar to the one presented in Sec. 5.4.2.

5.4.4 Tests in bounded domain

Dirichlet boundary conditions

Numerical parameters are the same as the ones presented in Sec. 5.4.1 for Dirichlet boundary conditions ($\beta = 0$ and $\gamma = 1$ in Eq. 5.39). We consider the aligned transfer method with the three boundary versions (*Methods 1, 2* and *3*). Their impact on the residual elimination in a multigrid V-cycle is presented in Fig. 5.19 for $K_{b\|} = 1$ and $K_{b\|} = 10^6$. In both cases, the *Method 3* shows the best performance, being the least influenced by $K_{b\|}$.

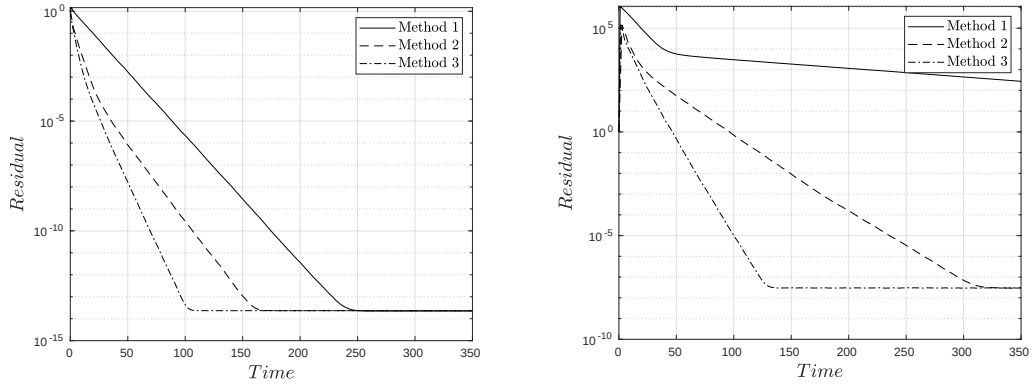


Figure 5.19: Time evolution of the residual with the present method for $K_{b\|} = 1$ (a) and $K_{b\|} = 10^6$ (b). The 3 methods are used near the boundary in the transfer matrix with Dirichlet boundary condition. Results for level 7 in the V-cycle. Total time has been limited to $350s$.

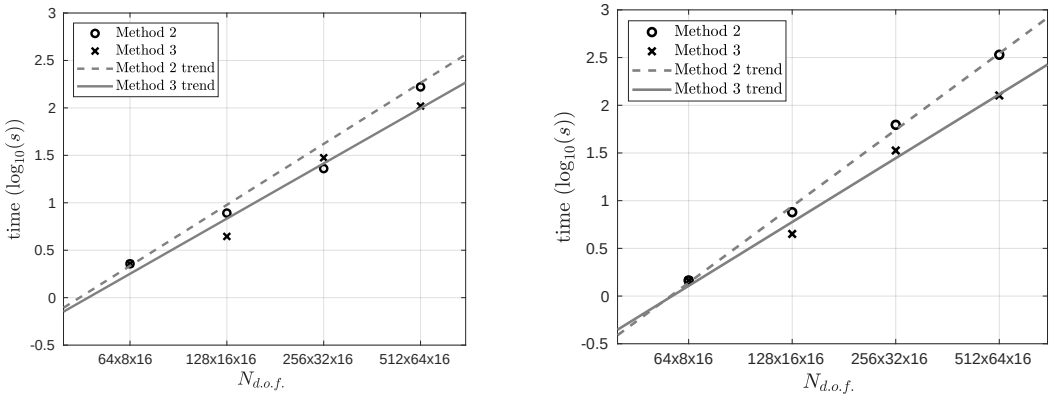
Time trends are also tested for a increasing number of unknowns and for *Methods 2* and *3*, Fig. 5.20. The results confirm those presented in Fig. 5.19. The *Method 3* seems to be best.

Neumann boundary condition

The same tests have been carried out with Neumann boundary condition (Fig. 5.21). The *Method 1* does not converge to the solution and results have been omitted.

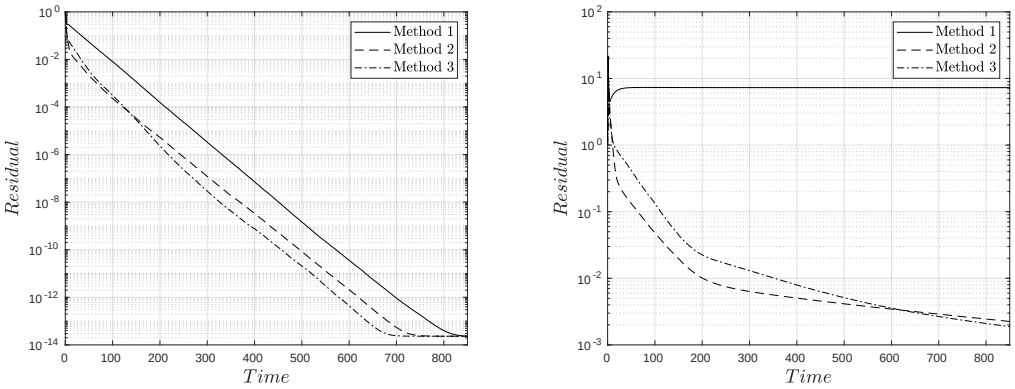
For low parallel diffusion, $K_{b\|} = 10^1$, Fig. 5.21a, the *Methods 2* and *3* shows a slower convergence than for Dirichlet boundary conditions and the *Method 3* provides however a better efficiency to converge the residual than the *Method 2*.

For high parallel diffusion, $K_{b\|} = 10^6$, Fig. 5.21 b, the residual term converges slowly that leads to a relative high timing interval ($\leq$ 10 mins.). Then, none of the methods presented in this section offers an optimal way to converge to the solution with a Neumann boundary condition and a high parallel diffusion.



(a) **(b)**

Figure 5.20: Time to reach the minimal residual term solving the linear system Eq. 5.2 for for (a) $K_{b\|} = 1$ and (b) $K_{b\|} = 10^6$ using the transfer *Methods 2* and *3* for the Present scheme with Dirichlet boundary conditions.



(a) **(b)**

Figure 5.21: Time to reach the minimal residual solving the linear system Eq. 5.2 for (a) $K_{b\|} = 1$ and (b) $K_{b\|} = 10^6$. The transfer *Methods 2* and *3* are used for the present scheme with Neumann boundary condition. *Method 1* does not convergence with Neumann boundary condition.

Robin boundary condition

As introduced in Sec. 3.48, this boundary condition is relevant to fusion plasma application since it is involved in several conservation equations [TGT+09, GTB+17, TGT+10].

Although results obtained with a Neumann boundary condition show a slow convergence, the combination with a Dirichlet boundary condition could make the system easier to solve, and leads to the convergence over a competitive time with respect to other methods. Fig. 5.22 shows the time evolution of the residual value for $K_{b\|} = 1$ and

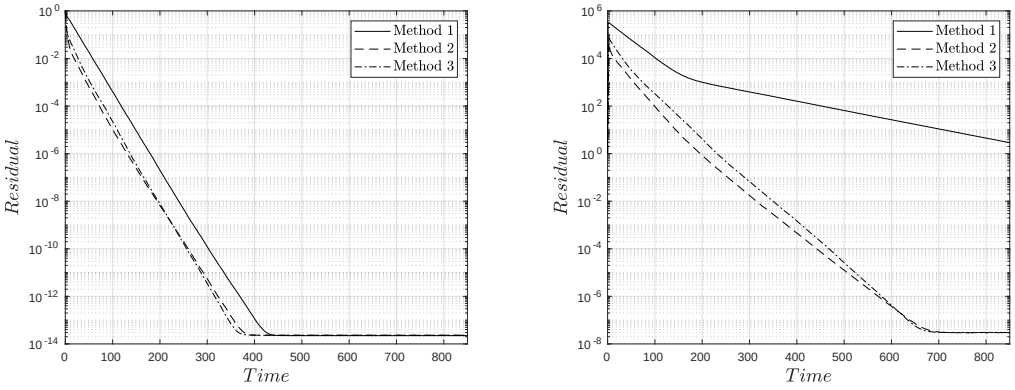


Figure 5.22: Time to reach the minimal residual term solving the linear system Eq. 5.2 for (a) $K_{b\|} = 1$ and (b) $K_{b\|} = 10^6$. The transfer *Methods 2* and *3* are used for the present scheme with a Robin boundary condition ($\gamma = 1$ and $\beta = 1$).

$K_{b\|} = 10^6$. The convergence is slower than for the Dirichlet boundary condition, but it is, however, faster than for Neumann boundary condition. The *Methods 2* and *3* are thus competitive to solve such kind of anisotropic problem with a Robin condition and a high parallel diffusion.

In conclusion, results for the aligned multigrid algorithm seems to be less competitive in the bounded domain than in a periodic one. It is especially true with a Neuman boundary condition. However, some important characteristics have emerged in these tests. The transfer adaptations *Methods 2* and *3* have shown to be convergent at low and high parallel diffusions, requiring only a reduced memory for solving large linear systems compared to the Matlab direct solver. All these good properties allow us to consider the multigrid algorithm presented in this Chapter as a possibly good preconditioner of an iterative solver.

5.5 The preconditioned generalized minimum residual method (GMRES)

5.5.1 The iterative solver

Multigrid algorithms are largely used as standalone solvers. In order to optimize the convergence towards zero of the residual at each iteration, the GMRES method introduced by Saad and Schultz [SS86] is considered here as a solver of the linear system 5.2. The multigrid algorithm can then be used as a preconditioner to improve the convergence properties of the GMRES solution. This combination is discussed as a Krylov acceleration method [TS01].

Since the goal of this section is only to test the GMRES algorithm with the multigrid routine as left preconditioning of the linear system, only the GMRES algorithm is described here. Full details about the Krylov subspace methods to solve linear systems can be consulted in [Saa03]. Essentially, the GMRES method is a generalization of the Paige and Saunders MINRES method [PS75] for nonsymmetric linear systems, both methods being based on the Lanczos method to solve the eigenvalue problem for a N×N matrix A and their relation with the conjugate gradient method. It is however suited for the solution on non-symmetric linear problems, and it is considered here in the view of using it for elliptic problems in bounded domains.

The GMRES is based on a projection on Krylov subspaces. For A the discrete operator matrix and v a vector, the projection on Krylov subspace leads to:

$$\mathcal{K}_m(A,u) \equiv span\{v, Av, A^2v, ..., A^{m-1}v\} \tag{5.57}$$

m being the matrix size. $\forall$ vector $v \in \mathbb{R}^N$ it has an exact projection $\mathcal{K} = \mathcal{K}_m$ in the m^{th} Krylov subspace. The basic GMRES algorithm computes the orthogonal projection of the residual $r = b - Au$ in $\mathcal{K}_m(r)$, which minimizes the distance between r and the basis vectors of $\mathcal{K}_m$. The basis is built establishing $v_1 = r_0/\|r_0\|$ and $\tilde{v}_k = Av_{k-1}$. The orthonormal basis is obtained by computing and applying the Gram-Schmidt method, leading to:

$$v_i^{'} = \tilde{v}_i - \sum_{j=1}^{i-1} \frac{\langle \tilde{v}_i, \tilde{v}_j \rangle}{\langle \tilde{v}_j, \tilde{v}_j \rangle} \tilde{v}_j \quad ; \quad v_i = \frac{v_i^{'}}{\|v_i^{'}\|} \tag{5.58}$$

At this point, the GMRES approximation considers the unique vector which minimizes the norm $\|r_0 - \alpha_k A v_k\|^2$, the approximated solution being given by:

$$u = u_0 + \sum_{k=1}^{m} \alpha_k v_k, \tag{5.59}$$

Then, depending on the dimension m of the Krylov subspace considered, the minimization of the norm is not expensive since it requires the solution of a $(m+1) \times m$ system, which is usually small.

The algorithm used here is given in 4.

5.5.2 Preconditioned GMRES

In comparison with direct solvers, iterative methods show better performance in terms of memory requirement and computing time for solving very-large non-symmetric linear systems. However, iterative methods suffer a lack of robustness due to the wide variation of time during the convergence solving nonsymmetric linear systems. The use of a

Data: $r_0 = b - Au_0$; $v_1 := r_0/\|r_0\|$
Define a $(m+1) \times m$ matrix $H_m = \{\alpha_{ij}\}_{1 \leq i \leq m+1, 1 \leq j \leq m}$. Set $\overline{H}_m = 0$;
for *j=1,2,...,m* **do**
  $w_j := Av_j$;
  for *i=1,2,...,j* **do**
    $\alpha_{ij} := \langle w_j, v_i \rangle$;
    $w_j := w_j - \alpha_{ij} v_i$;
  end
  $\alpha_{j+1,j} = \|w_j\|^2$;
  if $\alpha_{j+1,j} = 0$ **then**
    $m := j$ and go to last line
  end
  $v_{j+1} = w_j/\alpha_{j+1,j}$;
end
y_m=min$\|(\|r_0\|e_1) - \overline{H}_m v_m\|^2$ and $u_m = u_0 + V_m v_m$;

Algorithm 4: GMRES algorithm from Saad [Saa03]

preconditioned matrix can accelerate the convergence of iterative algorithms, improving the robustness and efficacy by reducing drastically the number of iterations and the computing time for the convergence of the residual. Preconditioning is just a reformulation of the original linear system to an equivalent one (i.e. with the same solution):

$$M^{-1}Au = M^{-1}b \tag{5.60}$$

where M is the *preconditioner* matrix of the linear system. The position of the preconditioner defines Eq. 5.60 as a *left-preconditioned* linear system. The condition of an effective precondition matrix M, apart of maintaining the same system solution, is to increment the iterative algorithm performance and make it easily inverted. Details on preconditioning are found in [Saa03, Dem97].

In this section, the preconditioned GMRES iterative method is studied. Then, 3 left-preconditioners are considered here:

- The GMRES with ILU0 preconditioner. The incomplete LU factorization (ILU0) is extensively used due to its simple implementation, and its good performance in terms of memory and time [RAKKSRG13, MAK03]. The choice as GMRES preconditioner is given by the acceptable quality of the results compared with those of other ILU decompositions as showed by Ghai et al. [GLJ16]. Details on ILU0 as a preconditioner are described in [Saa03].

- The GMRES with a V-cycle multigrid with Gauss-Seidel smoother as a preconditioner. Two versions are compared here varying the Laplacian discretization

and the transfer method: the Günter non-aligned Laplacian with the non-aligned interpolation ($x-$ direction reduction) as the transfer method (denoted as GMRES+NA MG(GS)), and the present Scheme aligned Laplacian with the aligned transfer method (denoted as GMRES+A MG(GS)).

- The GMRES with a V-cycle multigrid with GMRES smoother as a preconditioner. Since the Gauss-Seidel smoothing requires an inversion of a matrix at all iterations (which can be penalizing for the computing time) and due to the complexity to parallelize the calculus, a variation of the V-cycle multigrid algorithm is also introduced here using the GMRES also as a smoother after each transfer stage. Three multigrid versions are also compared here: the GMRES+NA MG(GMRES) with the Günter's scheme and the non-aligned transfer, GMRES+A MG(GMRES) with the present scheme as Laplacian discretization, and the GMRES+A MGG(GMRES) with aligned transfer and Günter's scheme as the Laplacian discretization.

Two test cases are tested in this section solving the linear system $Au = b$ in the 3D domain defined in Sec. 5.4.1. The first test case is done using just a random signal as:

$$S_a(i,j) = X(i,j) \tag{5.61}$$

where $X(i,j) \in U(]0,1[)$ is a random $\mathbb{R}$ number between 0 and 1. The source term S_a is thus a signal with the maximal number of modes contained in the x-y plane.

In order to model a characteristic signal of highly anisotropic flows with a moderate number of modes, the following test case is also proposed:

$$S_a(x,y) = C_1 + C_2 \sum_{1}^{20} sin(m_x x + m_y y) + C_3 \sum_{1}^{50} sin(m_{x2} x + m_{y2} y) \tag{5.62}$$

with $C_1 = C_2 = C_3 = 1$, $m_x = X_1$, $m_y = 7X_1$, $m_{x2} = X_2$ $m_{y2} = 7X_3$ and where X_1 is a set of 20 random integer numbers $\in U([1,5])$, and X_2, X_3 a set of 50 random integer numbers $\in U([-5,5])$ with $X_2 \neq X_3$. Then, the first sum of Eq. 5.62 generates aligned modes with a pitch angle $\alpha = atan(1/7)$, and the second sum an oriented but non aligned set of modes since $X_2 \neq X_3$.

5.5.3 Comparative tests

Periodic domain

The first test case considers the signal Eq. 5.61 as a r.h.s. of the linear system related to the resolution of the Helmholtz equation Eq. 5.1, as described in Sec. 5.4.1 ($K_{b\perp} = 0$). For a low parallel diffusion, $K_{b\parallel} = 1$, Fig. 5.23a, results show that all methods converge

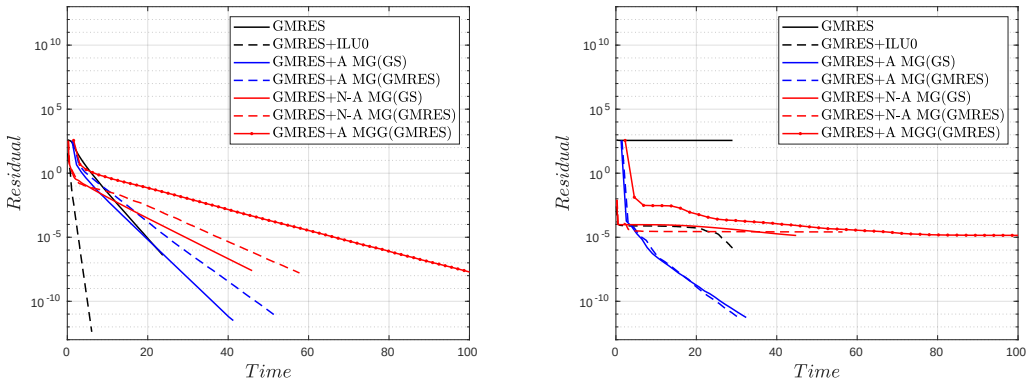


Figure 5.23: Time to converge the residual to a low value (10^{-9}). (a) $K_{b\|} = 1$; (b) $K_{b\|} = 10^6$ for the right hand side S_a defined by Eq. 5.61.

(residual going towards zero), the GMRES with the ILU0 preconditioner providing the best performance in terms of computing time.

For high parallel diffusion, $K_{b\|} = 10^6$, Fig. 5.23b, the GMRES does not converge during 200 cycles. None of the non aligned multigrid preconditioners presents a satisfactory convergence of the residual for 200 iterations. Since a part of the modes is solved during the first cycles, the algorithm does not converge to $Residual < 10^{-5}$. However, the GMRES with aligned transfer and the present scheme for the Laplacian discretization presents the best convergence rate, independently of the smoothing algorithm. The ILU(0) preconditioner with the GMRES presents a delayed convergence in comparison with the aligned multigrid preconditioners in this case.

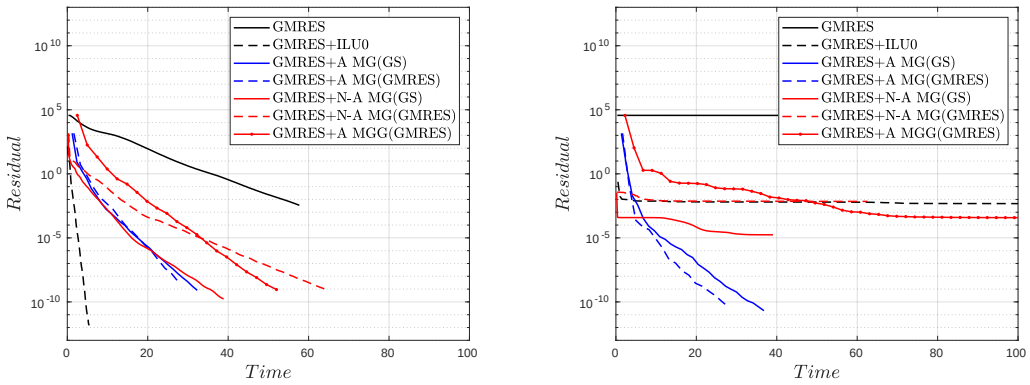


Figure 5.24: Time to converge the residual to a low value (10^{-9}). (a) $K_{b\|} = 1$; (b) $K_{b\|} = 10^6$ for the right hand side S_a defined by Eq.5.62.

The second test case considers the signal 5.62 as a r.h.s. of the linear system. For a low parallel diffusion, $K_{b\|} = 1$, Fig. 5.24a, results show comparable results to the previous test case. The GMRES with the ILU(0) clearly provides better results in terms of time to converge the residual. However, the method requires a larger amount of GMRES cycles than with the aligned multigrid with the present scheme for the Laplacian discretization, Fig. 5.25a. This can have a negative impact on the memory requirement.

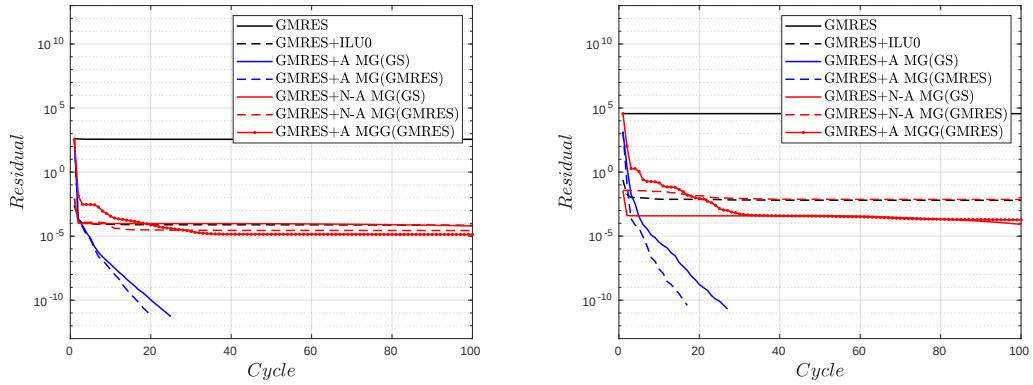


Figure 5.25: Number of GMRES cycles to converge the residual to a low value (10^{-9}). (a) $K_{b\|} = 1$ and (b) $K_{b\|} = 10^6$ for the right hand side S_a defined by Eq. 5.62 case.

For high parallel diffusion, $K_{b\|} = 10^6$, Fig. 5.24b, none of the non aligned multigrid preconditioners present a satisfactory residual elimination in 1000 GMRES cycles. Only the GMRES with aligned multigrid and aligned Laplacian methods achieves a rapid residual removing. In all tested cases, the GMRES with the ILU(0) precondition converges to a low final residual, but the computational time and the number of cycles seems to be influenced by the value of the parallel diffusion $K_{b\|}$.

***$K_{b\|}$* variation.** The Fig. 5.26 shows the necessary number of cycles and time to converge to a residual of $Residual = 10^{-9}$ dependinig on the value of the parallel diffusion $K_{b\|}$. Independently of the smoother method used, the aligned multigrid preconditioners are not dependent on the value of $K_{b\|}$. There is even a reduction of the number of cycles for high values of $K_{b\|}$ (it is caused by the nearly total elimination of the non aligned modes). In highly anisotropic cases (high values of the parallel diffusion) the results provided by the ILU(0) preconditioner are clearly less good than the one obtained with the GMRES+A MG.

Variation of the initial resolution. The negative impact shown above of a high anisotropy when using the GMRES-ILU(0), has also an impact on the computational time needed for solving the linear system depending on initial resolutions. The Fig. 5.27

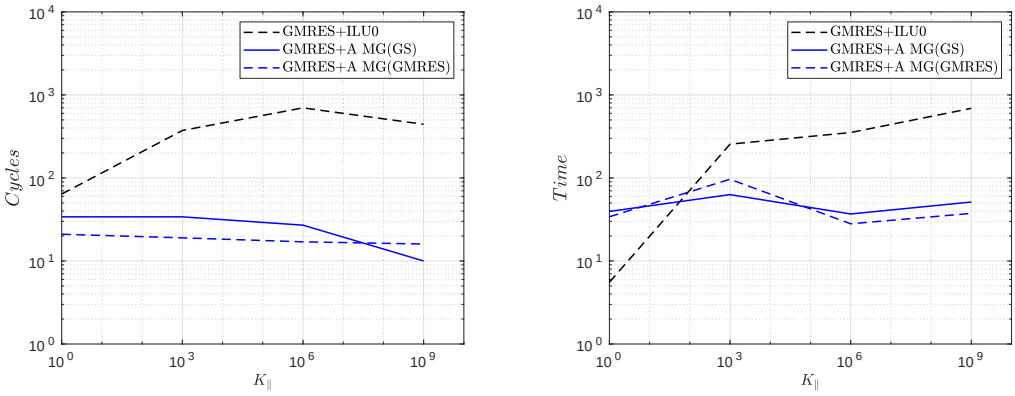


Figure 5.26: Number of GMRES cycles (a) and time (b) to converge the residual to a low value (10^{-9}) depending on the parallel diffusion $K_{b\parallel}$ and for different convergent GMRES preconditioners.

shows the time needed to converge to a low residual of 10^{-9} for $K_{b\parallel} = 1$ and $K_{b\parallel} = 10^6$ and depending on 3 initial resolutions $N_{dof} = 94208$, 524288, and 2097152. For low parallel diffusion, $K_{b\parallel} = 1$, the time increases nearly linearly with the resolution whatever the preconditioners. For high parallel diffusion, $K_{b\parallel} = 10^6$, the impact is very weak on the time evolution of the multigrid preconditioners, with about the same time to reach a residual of 10^{-9} for both values of $K_{b\parallel}$. However, the impact is strong for the GMRES-ILU(0) with a much deeper slope. Besides, for the highest resolution, the GMRES-ILU(0) is unusable due to the too large number of GMRES cycles needed (more than 700, $> 15.5GB$).

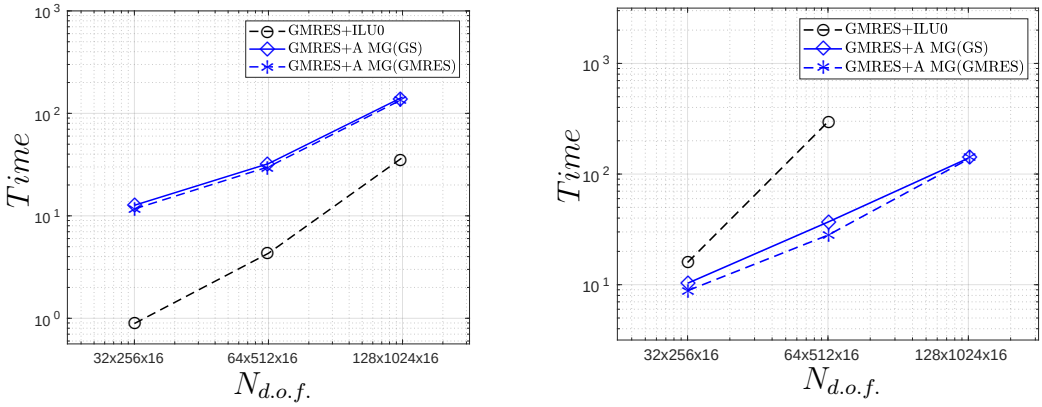
This test shows the efficiency of the multigrid preconditioners to deal with highly anisotropic problems and high resolutions.

Bounded domains

The boundary conditions in bounded domains involve important changes in the Laplacian discretization, and in the aligned transfer method.

Tests will be restricted to the most relevant cases, i.e. GMRES, GMRES+ILU(0), GMRES+A MD(GS) and GMRES+A MD(GMRES). The other methods tested did not converge in the periodic case.

Furthermore, due to the comparable results obtained in Sec. 5.4.4 between the transfer *Methods 2* and *3* for bounded problems, the multigrid with aligned transfer using the *Method 2* is also introduced as GMRES preconditioner, implemented with the Gauss-Seidel smoothing (identified as GMRES-A2 MD(GS)) and with the GMRES smoothing (GMRES-A2 MD(GMRES)). The identifiers GMRES+A MD(GS) and GMRES+A MD(GMRES) combines the *Method 3* as a transfer method near the boundaries.

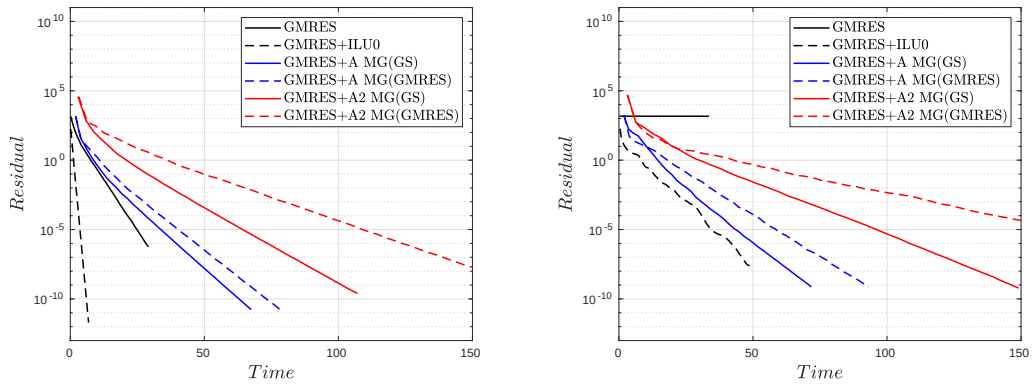


(a) **(b)**

Figure 5.27: Time to converge the residual to a low value (10^{-9}) depending on the initial number of degree of freedom.(a) $K_{b\|} = 1$; (b) $K_{b\|} = 10^6$.

In all the cases, boundary conditions (Dirichlet, Neumann or Robin) are imposed in the y-direction only, the x and z directions remaining periodic. This configuration is rather relevant for fusion plasma simulations at the edge of the reactor. A more relevant configuration would be to consider a bounded domain in the z-direction as well but is out of the scope of this thesis where we focus on high anisotropy flows in the 2D x-y plane.

Presented tests considers Eq. 5.62 as r.h.s. with the same number of random modes solving a $64 \times 512 \times 16$ system. We first consider Dirichlet boundary condition ($\gamma = 1$, $\beta = 0$ in Eq. 5.39).



(a) **(b)**

Figure 5.28: Time evolution of the residual calculated for 200 GMRES cycles when solving the linear system Eq. 5.2 with a Dirichlet boundary condition in the y-direction. (a) $K_{b\|} = 1$ and (b) $K_{b\|} = 10^6$.

Results are shown on Fig. 5.28 for $K_{b\|} = 1$ and $K_{b\|} = 10^6$. As expected, the GMRES without preconditioner lost convergence at high parallel diffusion for $K_{b\|} = 10^6$. The ILU(0) preconditioner always provides the best results whatever $K_{b\|}$. For a high parallel diffusion, results are comparable with those obtained with the multigrid preconditioners but do not allow to reach the same precision after 200 GMRES cycles.

Regarding to the boundary transfer method in the multigrid routine, the *Method 3* exhibits better performance than *Method 2* that reproduces the results shown in Fig. 5.19.

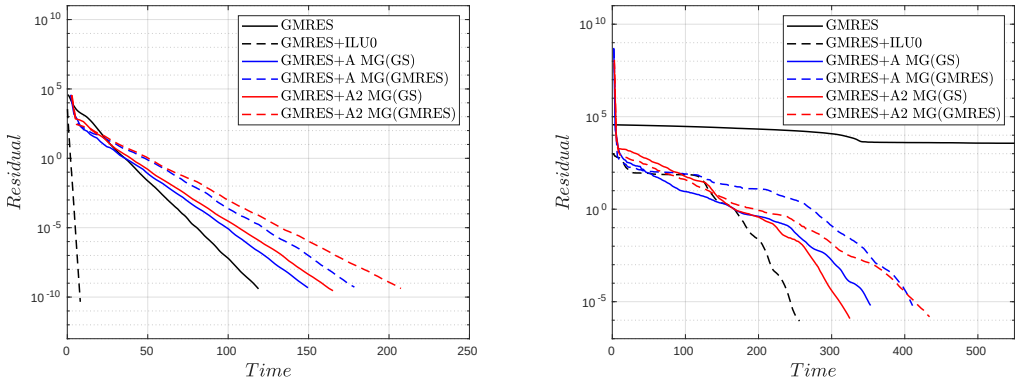


Figure 5.29: Time evolution of the residual calculated for 200 GMRES cycles when solving the linear system Eq. 5.2 with a Neumann boundary condition in the y-direction. (a) $K_{b\|} = 1$ and (b) $K_{b\|} = 10^6$.

We consider now a Neumann boundary condition in the y-direction, Fig.5.29. The results show relatively similar behaviour than for the Dirichlet case. The GMRES+ILU(0) method still presents the best performance whatever the parallel diffusion. The GMRES with multigrid preconditioners show a slower convergence and on contrary to the Dirichlet case do not provide a better accuracy after 200 GMRES cycles. Moreover, in contrast to the Dirichlet case, the use of *Methods 2* and *3* near the boundary in the transfer projection shows here comparable performance.

The aligned multigrid preconditioners method is not able to provide the same good properties than in the periodic configuration, with results comparable to the ILU(0) method for a high parallel diffusion.

5.5.4 Anisotropic Poisson equation in bounded domains

Tests on the Poisson's equation are here performed with a Robin boundary condition:

$$\begin{aligned} -\boldsymbol{\nabla}\cdot(\boldsymbol{\mathcal{K}}\cdot\boldsymbol{\nabla}))T &= S \qquad \text{on } \Omega \\ \frac{1}{R}\nabla_{b\|}T + T &= s \qquad \text{in } \Gamma \end{aligned} \tag{5.63}$$

The r.h.s. is the result of the analytical Laplacian considering $T = S_a$ Eq.5.61.

Considering $R = 1$, (same weight for the Dirichlet and the Neumann term), the Fig. 5.30, shows that all preconditioned GMRES solvers converge independently of $K_{b\|}$. The transfer *Method 3* gives better results than transfer *Method 2*, which is slower to converge. Here, ILU(0) and GMRES+A MG(GMERS) shows the same tendency for $K_{b\|} = 10^6$ in the convergence of the residual.

Testing now the solver performance with a more stringent Robin number, $R = 10^{-3}$, results on Fig. 5.31 show that at low parallel diffusion, $K_{b\|} = 1$, the multigrid preconditioner with the Gauss-Seidel algorithm provides the best performance. The same value of the residual is found that with the ILU(0) preconditioner but with about $\sim 40\%$ less time.

However, at high parallel diffusion, $K_{b\|} = 10^6$, the multigrid preconditioners need about the same time to converge to a residual of 10^{-9} when using *Method 3* near the boundary. Here, the ILU(0) preconditioner takes more time than the other multigrid preconditioners. In this case, the transfer *Method 2* improves the global performance compared with *Method 3*, reducing of about $\sim 40\%$ the computational time for the GS smoothing method.

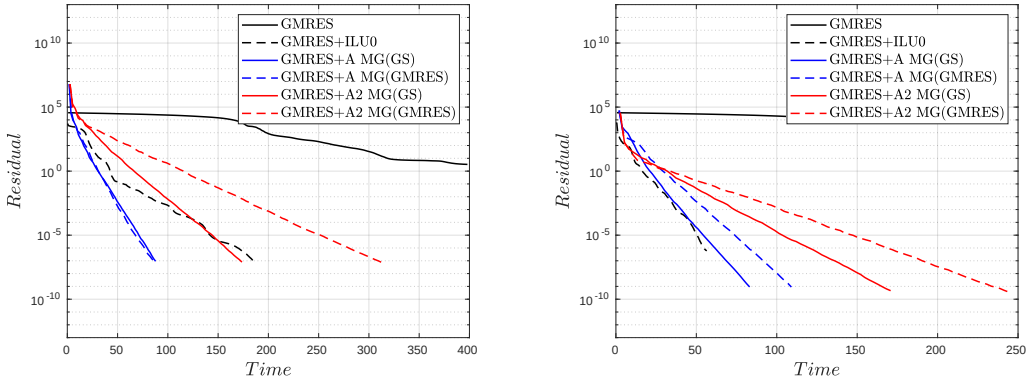


Figure 5.30: Time evolution of the residual when solving the Poisson equation with a Robin boundary condition in the y-direction with $R = 1$. (a) $K_{b\|} = 1$ and (b) $K_{b\|} = 10^6$.

The comparison in the number of GMRES cycles depending on the preconditioners is shown on Fig. 5.32 for $K_{b\|} = 10^6$ and for $R = 1$ and $R = 10^{-3}$. As shown in Sec. 5.5.3, the number of cycles is directly related to the total memory used for each preconditioned GMRES. Here, the definition of the boundary conditions has a huge impact on the global number of iterations needed. In both cases, the multigrid preconditioners need fewer cycles to converge the residual to 10^{-9} than the ILU(0) preconditioner.

The tests have shown that the configuration of the multigrid preconditioner depends on the nature of the boundary condition. Indeed, the transfer *Method 3* seems to provide

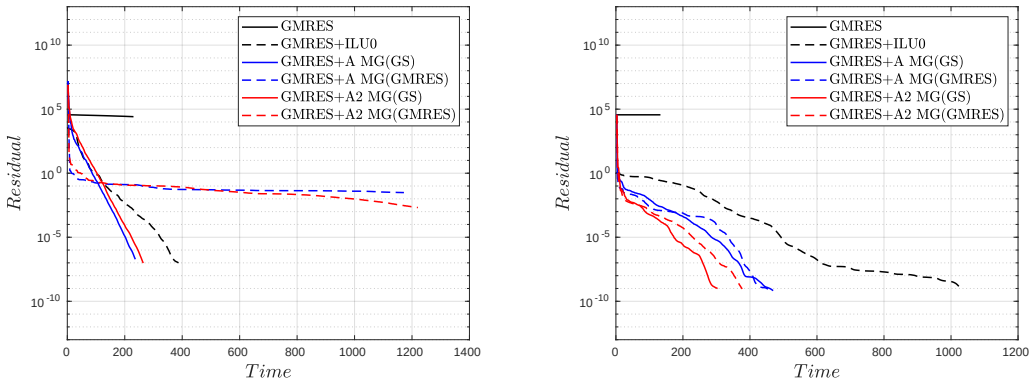


Figure 5.31: Time evolution of the residual when solving the Poisson equation with a Robin boundary condition in the y-direction with $R = 10^{-3}$. (a) $K_{b\|} = 1$ and (b) $K_{b\|} = 10^6$.

better results than *Method 2* when Dirichlet boundary condition is used whereas *Method 2* provides better results for ill-conditioned cases with $K_{b\|} = 10^6$ and R^{-3}.

Contrary to the tests performed in the bounded domain and for the Helmholtz equation, the multigrid preconditioners seem to be more stable than the ILU(0) preconditioner when solving the Poisson equation.

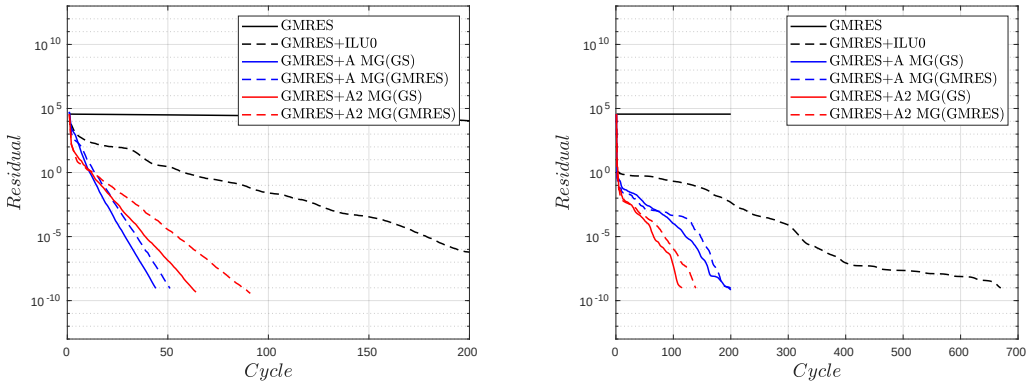


Figure 5.32: Number of cycles to converge the residual to 10^{-9} when solving the Poisson equation with a Robin boundary condition in the y-direction. (a) $K_{b\|} = 10^6$ and $R = 1$; (b) $K_{b\|} = 10^6$ and $R = 10^{-3}$.

5.6 Conclusion and perspectives

With the aim of obtaining an efficient algorithm for the resolution of highly anisotropic Helmoltz problems with a large number of unknowns, we have investigated a multigrid algorithm whose key component is a coarsening strategy based on identifying aligned fluctuations. This multigrid algorithm is implemented with standard V-cycles.

Its convergence has been tested on a bi-periodic anisotropic diffusion problem. Convergence is effectively obtained when the discretization is performed using aligned discretizations. Application of the multigrid algorithm to non-aligned discretizations is problematic, as convergence is not obtained when the problem is discretized using Günter's scheme. In case where convergence is obtained, the multigrid appears competitive: it shows favourable scaling in terms of computing time and memory requirements, and makes it preferable to direct inversion using the backlash solver of Matlab for large but reasonable grid sizes, especially in highly anisotropic diffusion cases.

The approach has been generalized to bounded domains, where the impact of 3 different strategies in coarsening the boundary has been assessed: in the first, boundary points are left out by the coarsening process, in the second, coarsening retains boundary points but the boundary points are not used in the prolongation step, whilst in the third boundary points are kept by the coarsening process and used in the prolongation step. Numerical experiments show the superiority of the third method in problems with Dirichlet and Neumann boundary conditions. Method 2 has lower but still comparable efficiency. The use of the first method however is largely detrimental to the multigrid algorithm, and should be avoided especially in cases where the parallel diffusivity is very large.

Building on these numerical experiments, multigrid acceleration is considered, i.e. the multigrid algorithm is proposed as preconditioner for a GMRES solution. This preconditioning is evaluated against ILU0-preconditioning on a series of anisotropic diffusion problems with periodic, Dirichlet, Neumann and Robin boundary conditions. Results show how the use of an aligned multigrid with an aligned Laplacian discretization is better than all the preconditioners and GMRES method in periodic domains, even for high $K_{b\|}$ values. The situation is different in bounded problems, where the ILU0 is often better in terms of performance than the proposed multigrid algorithm regardless of boundary condition and parallel diffusivity. The same evaluation is finally conducted for bounded Poisson problems: in this case, ILU0 preconditioning is more efficient, except in the demanding case with strong parallel diffusivity and in which the gradient has a large contribution to the Robin boundary conditions. This case corresponds to a badly ill-conditioned problem, and the multigrid algorithm is unequivocally superior.

Although the aligned multigrid solver presented in this work seems to be well adapted to highly anisotropic flows, there are still some points that could be improved and which have not been explored in this work. We can mention, for example, the V-cycle that could be modified to an F-cycle, see [Sha95], and which concentrates the transfer and

smoothing in the lowest levels (reduced grid, where the slow modes are found) and could thus lead to better convergence. In addition, the contrast observed between the good results of the aligned multigrid obtained in a periodic domain with respect to the results in a bounded domain suggests exploring other transfer methods near the boundary.

Chapter 6

Main conclusions and relevance with the implementation of aligned coordinates method in TOKAM3X

One of the primary goals of this PhD thesis was to investigate the efficiency of an aligned coordinates method for solving the fluid model implemented in TOKAM3X, compared to the currently implemented *Günter's* scheme. In the following, we list the benefits and requirements of such a scheme according to the results obtained during this work. Obviously, these conclusions are valuable beyond the fluid model of TOKAM3X and can be a benefit for any highly anisotropic diffusive fluid model implemented in any code based on finite differences.

6.1 On the Laplacian discretization in highly anisotropic diffusion

The scheme developed during this thesis (Chapter 3) has been designed for standard Cartesian grids, but using interpolations aligned along the parallel direction (defined as the direction of anisotropy). Based on the Support Operator Method (SOM) as the *Günter's* or *Stegmeir's* schemes, the present scheme maintains the self-adjointness property of the parallel diffusion operator at the discrete level.

Numerical tests based on manufactured solutions have shown that all features of the aligned methods mentioned in the literature are recovered, in particular, the fact that aligned methods allow to drastically reduce the number of mesh points with respect to non-aligned approaches since precision becomes N_y dependent only in aligned methods. This reduction becomes even more significant as the anisotropy is increased due to the fact that non-aligned modes are rapidly damped by the operator, so in this case, much fewer points in the x-direction (N_x) are required to get a good accuracy of the solution.

Let's notice that the new scheme brings, however, new features compared to the literature, which are crucial to get accurate solutions in most of the realistic applications:

- Contrarily to existing schemes of the literature based on aligned interpolations, the *present* scheme guarantees conservativity of the fluxes, not only along the parallel direction but also in the direction across the main diffusion direction, which had not been addressed in former papers.

- The proposed method has been also extended to obtain a conservative advection scheme. Since the fluxes along the parallel direction are obtained in the Control Volume (CV) limits in a commonly defined space, the balance of these flux in a given CV leads to the parallel advection term. This feature is not shown in *Stegmeir's* scheme.

- A method to deal with domain boundaries has been proposed, which is compatible with the aligned discretization adopted for inner nodes. This method provides much better accuracy than the classical approach based on ghost points which are usually extrapolated far away outside the domain along the anisotropy direction. In addition, the method proposed in this thesis remains compatible with the mesh reduction used in the inner domain.

In the TOKAM3X code, the discrete temperatures (Helmholtz-like equation) and vorticity (Poisson-like equation) anisotropic operators are discretized using the Günter's scheme with a usual grid of size $64 \times 512 \times 32$. The anisotropic diffusion being aligned along the radial direction (the z-direction in Cartesian grids), and non aligned in the toroidal-poloidal direction (the x-y plane in Cartesian grids), the magnetic surfaces are contained in this plane. Then, results shown in Fig. 3.15 in Sec. 3.5.3 are relevant. They show that for a grid of size 64×512, the *Günter's* scheme is better adapted than an aligned method as long as the anisotropy is weak but it exhibits a high numerical diffusion when the anisotropy becomes large, $K_{b\|} = 10^6$ in the proposed tests cases, these latest being much more relevant for fusion applications.

The results of this thesis have shown that a reorganization of the grid points can be a benefit to the implementation of an aligned approach. The Fig. 6.1 below shows 2D plots of the $\mathcal{H}^1$-error related to the test case detailed in Sec. 3.5.1 for $K_{b\|} = 10^6$, i.e. a highly anisotropic flow relevant for fusion applications. The red stars indicate the grid size 64×512 currently used to discretized the x-y-plane in TOKAM3X simulations, and the red dots indicate different resolutions in the x and y directions but keeping the same total number of degrees of freedom, i.e. $N_{d.o.f.} = 32768$. Using the aligned approach developed during the thesis, a grid size of 16×2048 provides a $\mathcal{H}^1$-error of about 10^{-1} error while a grid size of 8×4096 provides a $\mathcal{H}^1$-error of about 10^{-2}. This redistribution limits the number of parallel modes represented in the simulation, since non aligned modes with a large orientation with respect to $\mathbf{b}$ are limited to 16 and 8 points, respectively.

However, a higher resolution in the perpendicular dynamics (y-direction) will increase the precision in this direction, which very benefits for TOKAM3X simulations since the study of the perpendicular dynamics is one of the main objectives.

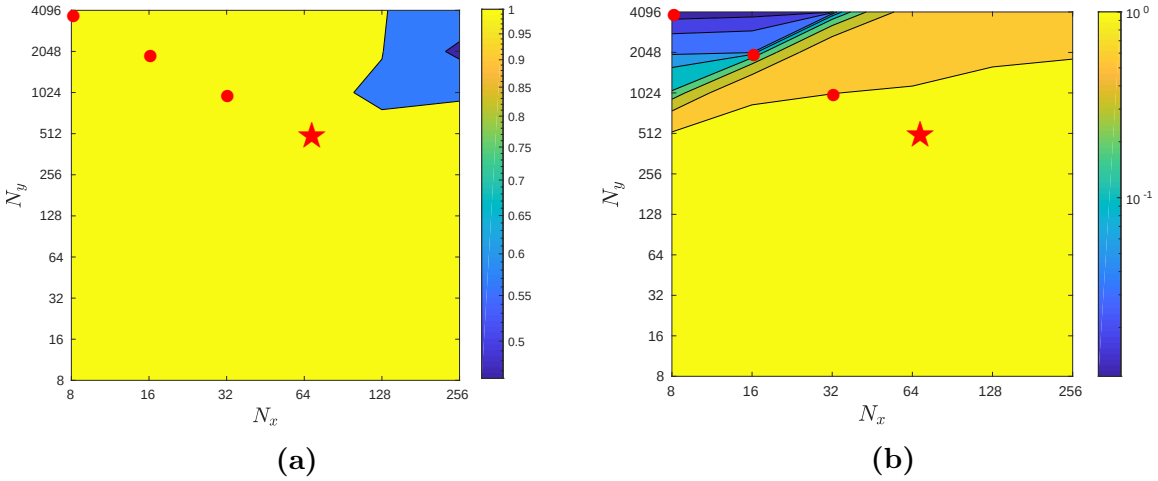


Figure 6.1: 2D plots of the $\mathcal{H}^1$-error in function of N_x and N_y. Bi-periodic computational domain with $K_{b\|} = 10^6$. (a) *Günter's method*; (b) *Present method*

6.2 On the field decomposition

Filtering methods can be an interesting perspective in order to eliminate the numerical diffusion caused by the aligned modes. It leads the discrete Laplacian operator untouched since the filtering application is based on a reformulation of Helmholtz equation where equality $\overline{T} = \overline{S}_a$ has been proven.

In Figs. 6.2 we compare the same filtering process (LA1 with $10,000$ *its*) solving the tests case detailed in Sec. 3.5.1 using the *Günter's* (a) and the *present* (b) schemes for $K_{b\|} = 10^6$. In both cases, the filtering method is able to eliminate the numerical pollution, showing the benefits in terms of precision given by the aligned method of the thesis. Since for the original TOKAM3X resolution (red star in Figs. 6.2), both methods exhibits the same $\mathcal{H}^1$-error of about 10^{-1} order. On contrary to the *Günter's* scheme, the *present* scheme exhibits a lower $\mathcal{H}^1$-error by doing a redistribution of the grid points but by keeping the same total number of points, $N_{dof} = 32,768$ (red dots in Figs. 6.2). For the *Günter's* scheme, the only way to increase the precision is to increase the total number of grid points and so the computing time and memory requirement.

In TOKAM3X, the filtering methods could be only applicable for the temperature equations, which can be recast as a Helmholtz equation. However, the method could not be applicable to solve the vorticity equation involving a Poisson equation.

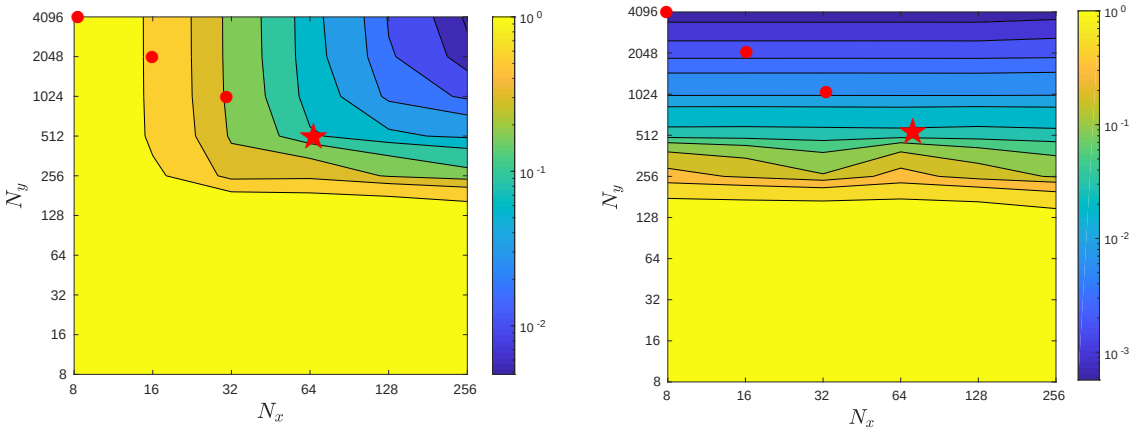


Figure 6.2: 2D plots of the $\mathcal{H}^1$-error in function of N_x and N_y. Bi-periodic computational domain with $K_{b\|} = 10^6$. (a) *Günter's method.*(b) *Present method*

6.3 On the iterative solvers

An iterative solver has been proven an optimal way to solve large linear systems in terms of computational time and memory. In Chapter 5 a multigrid algorithm has been specially developed for highly anisotropic elliptic problems, leading to an aligned treatment of the transfer methods. Combined with the aligned methods it is the better way to obtain a robust iterative solver. The multigrid adapted to the *Günter's* scheme has been proven to be inefficient in highly anisotropic flows.

The GMRES iterative solver has been tested with different preconditioners. The aligned multigrid (aligned transfer and Laplacian) preconditioner has been shown to provide the best results in the periodic computational domain. However, its efficiency has been shown to decrease in bounded domains when solving the Helmholtz equation (even if it remains competitive in terms of memory requirement). For the Poisson problem with a Robin boundary condition, the multigrid preconditioner provides the best results.

In conclusion, a GMRES iterative solver combined with the aligned multigrid preconditioner proposed in this thesis can be an interesting way to solve efficiently linear systems in TOKAM3X but only if an aligned approach is considered to discretize the Laplacian.

Bibliography

[Aav07] I. Aavatsmark. Multipoint flux approximation methods for quadrilateral grids. *9th International Forum on Reservoir Simulation*, 01 2007.

[ABBM94] I Aavatsmark, T Barkve, Ø Bøe, and T. Mannseth. Discretization on non-orthogonal, curvilinear grids for multi-phase flow. *4th European Conference on the Mathematics of Oil Recovery*, 06 1994.

[ABBM96] I. Aavatsmark, T. Barkve, Ø. Bøe, and T. Mannseth. Discretization on non-orthogonal, quadrilateral grids for inhomogeneous, anisotropic media. *J. Comp. Phys.*, 127(1):2 – 14, 1996.

[ADLT06] S. Amat, R. Donat, J. Liandrat, and J. C. Trillo. Analysis of a new nonlinear subdivision scheme. applications in image processing. *Found. Comput. Math.*, pages 193–225, 2006.

[AL05] S. Amat and J. Liandrat. On the stability of the pph nonlinear multiresolution. *Appl. Comput. Harm. Anal.*, 18:198–206, 2005.

[Ama08] S. Amat. A review on the piecewise polynomial harmonic interpolation. *Appl. Numer. Math.*, 58:1168–1185, 2008.

[Ara97] A. Arakawa. Computational design for long-term numerical integration of the equations of fluid motion: two-dimensional incompressible flow, part i. *J. Comp. Phys.*, 135:103–114, 1997.

[BBC+13] B. Bensiali, K. Bodi, G. Ciraolo, Ph. Ghendrih, and J. Liandrat. Comparison of different interpolation operators including nonlinear subdivision schemes in the simulation of particle trajectories. *J. Comp. Phys.*, 236:346–366, 2013.

[BHM00] W. Briggs, V. Henson, and S. McCormick. A multigrid tutorial, 2nd edition. *SIAM*, 2000.

[BP83] B. R. Baliga and S. V. Patankar. A control volume finite-element method for two-dimensional fluid flow and heat transfer. *Num. Heat Transfer*, 6:245–261, 1983.

[Bra65] S. I. Braginskii. Transport process in a plasma. *Rev. in Plasma Physics*, 1:205–311, 1965.

[BS92] I. Babuska and M. Suri. On locking and robustness in the finite element method. *J. Comp. Phys.*, 29(5):1261–1293, 1992.

[CC12] Y. Chen and B. Cockburn. Analysis of variable-degree hdg methods for convection-diffusion equations. part i: General nonconforming meshes. *IMA J Numer Anal*, 32:1267 – 1293, 2012.

[CC14] Y. Chen and B. Cockburn. Analysis of variable-degree hdg methods for convection-diffusion equations. part ii: Semimatching nonconforming meshes. *Mathematics of Computation*, 83:87 – 111, 2014.

[CDG08] B. Cockburn, B. Dong, and J. Guzmán. A superconvergent ldg-hybridizable galerkin method for second-order elliptic problems. *Math. Comput.*, 77:1887–1916, 10 2008.

[CGL09] B. Cockburn, J. Gopalakrishnan, and R. Lazarov. Unified hybridization of discontinuous galerkin, mixed, and continuous galerkin methods for second order elliptic problems. *SIAM Journal on Numerical Analysis*, 47(2):1319–1365, 2009.

[CKG+06] C.S. Chang, S. Ku, L. Greengard, H. Weitzner, D. Zorin, M. Adams, D. Keyes, G. D'Azevedo, S. Klasky, P. Worley, Y. Chen, S. Parker, J. Cummings, S. Ethier, T.S. Hahm, W.W. Lee, R. Samtaney, D. Stotler, F. Hinton, Z. Lin, Y. Nishimura, and Cpes Team. Integrated particle simulation of neoclassical and turbulence physics in the tokamak pedestal/edge region using xgc. *IAEA fusion energy conference*, 01 2006.

[CKS91] S. Cowley, R. Kulsrud, and R. Sudan. Considerations of ion-temperature-gradient-driven turbulence. *Phys. Fluids B*, 3(10):2767–2782, 1991.

[CKT+17] C.S. Chang, Seung-Hoe Ku, George Tynan, R Hager, R Churchill, Istvan Cziegler, M Greenwald, A E. Hubbard, and J W. Hughes. A fast l-h bifurcation dynamics in a tokamak edge plasma gyrokinetic simulation. *APS Division of Plasma Physics Meeting 2017*, 2017.

[Col15] C. Colin. Turbulent transport modelling in the edge plasma of tokamaks: Verification, validation, simulation and synthetic diagnostics. *PhD thesis, Aix-Marseille University*, 2015.

[CS97] B. Cockburn and C.W. Shu. The local discontinuous galerkin method for time-dependent convection-diffusion systems. *SIAM Journal on Numerical Analysis*, 35:2440–2463, 1997.

[Dav01] P. A. Davidson. An introduction to magnetohydrodynamics. *Cambridge Univ. Press*, 2001.

[Dav04] Timothy A. Davis. Algorithm 832: UMFPACK v4.3—an Unsymmetric-pattern Multifrontal Method. *ACM Trans. Math. Softw.*, 30(2):196–199, June 2004.

[dCNC11] D. del Castillo-Negrete and L. Chacon. Local and nonlocal parallel heat transport in general magnetic fields. *Physical review letters*, 106:195004, 05 2011.

[DDL+12] P. Degond, F. Deluzet, A. Lozinski, J. Narski, and C. Negulescu. Duality-based asymptotic-preserving method for highly anisotropic diffusion equations. *Commun. Math. Sci.*, 10(1):1–31, 2012.

[DDN10] P. Degong, F. Deluzet, and C. Negulescu. An asymptotic preserving scheme for anisotropic elliptic problems. *Multiscale Model. Simul.*, 8:645–666, 2010.

[Dem97] J. Demmel. Applied numerical linear algebra. *SIAM*, 1997.

[DG83] R. L. Dewar and A. H. Glasser. Ballooning mode spectrum in general toroidal systems. *Phys. of Fluids*, 26(10):3038–3052, 1983.

[DLNN12] P. Degond, A. Lozinski, J. Narski, and C. Negulescu. An asymptotic-preserving method for highly anisotropic elliptic equations based on a micro–macro decomposition. *J. Comp. Phys.*, 231:2724–2740, 2012.

[Dro14] J. Droniou. Finite volume schemes for diffusion equations: Introduction to and review of modern methods. *Mathematical Models and Methods in Applied Sciences*, 24(08):1575–1619, 2014.

[EHH+11] R. Eymard, G. Henry, R. Herbin, F. Hubert, R. Klofkorn, and G. Manzini. 3d benchmark on discretization schemes for anisotropic diffusion problems on general grids. *Proceedings of Finite Volume for complex applications VU*, 4, 2011.

[FKX13] X. Feng, O. Karakashian, and Y. Xing. Recent developements in discontinous galerkin finite element methods for partial differential equations. *Springer*, 157, 2013.

[Fre07] J. P. Freidberg. Plasma physics and fusion energy. *Cambridge Univ. Press*, 2007.

[FT96] J. Ferguson and I. W. Turner. A control volume finite element numerical simulation of the drying of spruce. *J. Comp. Phys.*, 125(79):59–70, 1996.

[Gar01] X. Garbet. Instabilities, turbulence et transport dans un plasma magnetisé. *Thèse d'habilitation à diriger des recherches. Univ. Provence Aix-Marseille I*, 2001.

[GB14] X. Garbet and P. Beyer. Physique et technologies des plasmas de fusion par confinement magnétique. *Master physique Sciences de la Fusion*, 2014.

[GBC+18] G. Giorgiani, H. Bufferand, G. Ciraolo, P. Ghendrih, F. Schwander, E. Serre, and P. Tamain. A hybrid discontinuous galerkin method for tokamak edge plasma simulations in global realistic geometry. *J. Comp. Phys.*, 374:515 – 532, 2018.

[GHHK18] M. Gander, L. Halpern, F. Hubert, and S. Krell. Optimized schwarz methods for anisotropic diffusion with discrete duality finite volume discretizations. *HAL*, 2018.

[GLJ16] A. Ghai, C. Lu, and X. Jiao. A comparison of preconditioned krylov subspace methods for nonsymmetric linear systems. *Num. Lin. Alg. with Appl.*, 07 2016.

[GR07] R. J. Goldston and P. H. Rutherford. Introduction to plasma physics. *Cambridge Univ. Press*, 2007.

[GTB+17] D. Galassi, P. Tamain, H. Bufferand, G. Ciraolo, Ph. Ghendrih, C. Baudoin, C. Colin, N. Fedorczack, N. Nace, and E. Serre. Drive of parallel flows by turbulence and large-scale e×b transverse transport in divertor geometry. *Nucl. Fusion*, 57(3), 2017.

[GTW+88] M. Greenwald, J.L. Terry, S.M. Wolfe, S. Ejima, M.G. Bell, S.M. Kaye, and G.H. Neilson. A new look at density limits in tokamaks. *Nucl. Fusion*, 28(12):2199–2207, 1988.

[GYKL05] S. Günter, Q. Yu, J. Krüger, and K. Lackner. Modelling of heat transport in magnetised plasmas using non-aligned coordinates. *J. Comp. Phys.*, 209:354–370, 2005.

[Ham62] S. Hamada. Hydromagnetic equilibria and their proper coordinates. *Nucl. Fusion*, 2:23–37, 1962.

[HBD+93] G.W. Hammett, M.A. Beer, W. Dorland, S.C. Cowley, and S.A. Smith. Developments in the gyrofluid approach to tokamak turbulence simulations. *Plasma Phys. and Con. Fusion*, 35:973–985, 1993.

[Her00] F. Hermeline. A finite volume method for the approximation of diffusion operators on distorted meshes. *J. Comp. Phys.*, 160(2):481 – 499, 2000.

[Her03] F. Hermeline. Approximation of diffusion operators with discontinuous coefficients on distorted meshes. *Comp. Meth. in Appl. Mech. and Eng.*, 192:1939–1959, 04 2003.

[HH08] R. Herbin and F. Hubert. Benchmark on discretization schemes for anisotropic diffusion problems on general grids. *Finite volumes for complex applications*, 5:659–692, 2008.

[HO13] F. Hariri and M. Ottaviani. A flux-coordinate independent field-aligned approach to plasma turbulence simulations. *Comp. Phys. Comm.*, 184:2419 – 2429, 2013.

[Jin12] S. Jin. Asymptotic preserving (ap) schemes for multiscale kinetic and hyperbolic equations: a review. *Riv. Mat. Univ. Parma*, 3:177–216, 2012.

[JMA13] P. Jacq, P.-H. Maire, and R. Abgrall. A high-order cell-centered finite volume scheme forsimulating three dimensional anisotropic diffusionequations on unstructured grids. *IMA Journal of Numerical Analysis*, 26, 2013.

[JT03] A. P. Jayantha and Ian W. Turner. Generalized finite volume strategies for simuating transport in strongly orthotropic porous media. *Australian & New Zealand Industrial and Applied Mathematics Journal*, 44(1):C1–C21, 2003.

[JT05] A. P. Jayantha and Ian W. Turner. A Second Order Control-Volume Finite-Element Least-Squares Strategy For Simulation Diffusion In Strongly Anisotropic Media. *J. Comp. Math.*, 23(1):1–16, 2005.

[KR05] Runhild A. Klausen and Thomas F. Russell. Relationships among some locally conservative discretization methods which handle discontinuous coefficients. *Comp. Geo.*, 8(4):341, 2005.

[Law57] J. D. Lawson. Some criteria for a power producing thermonuclear reactor. *Proc. Phys. Soc. (London)*, 70-1:6–10, 1957.

[LMJ87] E. W. Larsen, J. E. Morel, and W. F. Miller Jr. Asymptotic solutions of numerical transport problems in optically thick, diffusive regimes. *J. Comp. Phys.*, 69:283–324, 1987.

[LMS14] K. Lipnikov, G. Manzini, and M. Shashkov. Mimetic finite difference method. *J. Comp. Phys.*, 257:1163–1227, 2014.

[LOC94] X.D. Liu, S. Osher, and T. Chan. Weighted essentially non-oscillatory schemes. *J. Comp. Phys.*, 115(1):200 – 212, 1994.

[LP05] C. Le Potier. Finite volume monotone scheme for highly anisotropic diffusion operators on unstructured triangular meshes. *Comptes Rendus. Mathématique. Académie des Sciences, Paris*, 12, 12 2005.

[LPHO12] C. Le Potier and T. Hai Ong. A cell-centered scheme for heterogeneous anisotropic diffusion problems on general meshes. *Int. J. on Finite Volumes*, 8, 05 2012.

[MAK03] R.C Mittal and A.H Al-Kurdi. An efficient method for constructing an ilu preconditioner for solving large sparse nonsymmetric linear systems by the gmres method. *Comp. & Math. with Applications*, 45(10):1757 – 1772, 2003.

[MB11] P.H. Maire and J. Breil. A high-order finite volume cell-centered scheme for anisotropic diffusion on two-dimensional unstructured grids. *HAL*, 11:76–15, 2011.

[MB12] P.H. Maire and J. Breil. A nominally second-order accurate finite volume cell-centered scheme for anisotropic diffusion on two-dimensional unstructured grids. *J. Comp. Phys.*, 231(5):2259 – 2299, 2012.

[MN12] A. Mentrelli and C. Negulescu. Asymptotic-preserving scheme for highly anisotropic non-linear diffusion equations. *J. Comp. Phys.*, 231:8229–8245, 2012.

[MRS98] J. E. Morel, R. M. Roberts, and M. Shashkov. A local support-operators diffusion discretization scheme for quadrilateral r-z meshes. *J. Comp. Phys.*, 144:17–51, 1998.

[MS08] L. G. Margolin and M. Shashkov. Finite volume methods and the equations of finite scale: A mimetic approach. *Int. J. Num. Meth. Fluids*, 56:991–1002, 2008.

[MSS00] L. G. Margolin, M. Shashkov, and Piotr K. Smolarkiewicz. A discrete operator calculus for finite difference approximations. *Comp. Methods Appl. Mech. Engrg.*, 187:365–383, 2000.

[NO14] J. Narski and M. Ottaviani. Asymptotic preserving scheme for strongly anisotropic parabolic equations for arbitrary anisotropy direction. *J. Comp. Phys.*, 185:3189–3203, 2014.

[NPC09] N.C. Nguyen, J. Peraire, and B. Cockburn. An implicit high-order hybridizable discontinuous galerkin method for linear convection–diffusion equations. *J. Comp. Phys.*, 228(9):3232 – 3254, 2009.

[NPC11] N.C. Nguyen, J. Peraire, and B. Cockburn. An implicit high-order hybridizable discontinuous galerkin method for the incompressible navier–stokes equations. *J. Comp. Phys.*, 230(4):1147 – 1170, 2011.

[Nyq28] H. Nyquist. Certain topics in telegraph transmission theory. *American Telephone and Telegraph Co. AIEE Winter Convention*, 1928.

[org19] ITER organization. Plasma heating. *ITER.org*, 2019.

[Ott11] M. Ottaviani. An alternative approach to field-aligned coordinates for plasma turbulence simulations. *Phys. Letters*, 375:1677 – 1685, 2011.

[PKFC+17] D. Prisiazhniuk, A. Krämer-Flecken, G. D. Conway1, T. Happel, A. Lebschy, P. Manz, V. Nikolaeva, U. Stroth, and ASDEX Upgrade Team. Magnetic field pitch angle and perpendicular velocity measurements from multi-point time-delay estimation of poloidal correlation reflectometry. *Plasma Phys. and Con. Fusion*, 59(2), 2017.

[PM62] Daniel P. Petersen and David Middleton. Sampling and reconstruction of wave-number-limited functions in n-dimensional euclidean spaces. *Information and Control*, 5(4):279 – 323, 1962.

[PP08] J. Peraire and P. Persson. The compact discontinuous galerkin (cdg) method for elliptic problems. *SIAM Journal on Scientific Computing*, 30(4):1806–1824, 2008.

[PS75] C. Paige and M. Saunders. Solution of sparse indefinite systems of linear equations. *SIAM J. Num. Analysis*, 12(4):617–629, 1975.

[PT99] Patrick Perré and Ian W. Turner. Transpore: A generic heat and mass transfer computational model for understanding and visualizing the drying of porous media. *Drying Technology*, 17(7-8):1273–1289, 1999.

[RAKKSRG13] R R. Akhunov, Sergei Kuksenko, V K. Salov, and T R. Gazizov. Optimization of the ilu(0) factorization algorithm with the use of compressed sparse row format. *J. of Math. Sc.*, 191, 05 2013.

[RT65] K. V. Roberts and J.B. Taylor. Gravitational resistive instability of an incompressible plasma in a sheared magnetic field. *Phys. of Fluids*, 8(2):315–322, 1965.

[Saa03] Y. Saad. *Iterative Methods for Sparse Linear Systems*. Society for Industrial and Applied Mathematics, second edition, 2003.

[SCM+16] A. Stegmeir, D. Coster, O. Maj, K. Hallatschek, and K. Lackner. The field line map approach for simulations of magnetically confined plasmas. *Comp. Phys. Comm.*, 198:139–153, 2016.

[Sco01] B. Scott. Shifted metric procedure for flux tube treatments of toroidal geometry: Avoiding grid deformation. *Phys. Plasmas*, 8(2):447–458, 2001.

[SH07] P. Sharma and G.W. Hammett. Preserving monotonicity in anisotropic diffusion. *J. Comp. Phys.*, 227:123–142, 2007.

[Sha48] C. E. Shannon. A mathematical theory of communication. *The Bell System Technical Journal*, 27:379–423, 623–656, 1948.

[Sha95] V.V. Shaidurov. Multigrid methods for finite elements. *Springer*, 1995.

[SM90] P. Stangeby and G.M. McCracken. Plasma boundary phenomena in tokamaks. *Nucl. Fusion*, 30(7):1225–1379, 1990.

[SMC+17] A. Stegmeir, O. Maj, D. Coster, K. Lackner, M. Held, and M. Wiesenberger. Advances in the flux-coordinate independent approach. *Comp. Phys. Comm.*, 213:111–121, 2017.

[SS86] Y. Saad and M. Schultz. GMRES: A generalized minimal residual algorithm for solving nonsymmetric linear systems. *SIAM Journal on Scientific and Statistical Computing*, 7(3):856–869, 1986.

[SS94] M. Shashkov and S. Steinberg. Support-operator finite-difference algorithms for general elliptic problems. *J. Comp. Phys.*, 118:131–151, 1994.

[Sta00] P. Stangeby. The plasma boundary of magnetic fusion devides. *IOP Editors*, 2000.

[Tam07] P. Tamain. Etude des flux de matière dans le plasma de bord des tokamaks: Alimentation, transport et turbulence. *Thèse, Univ. Provence Aix-Marseille I*, 2007.

[TBC+16] P. Tamain, H. Bufferand, G. Ciraolo, C. Colin, D. Galassi, Ph. Ghendrih, F. Schwander, and E. Serre. The TOKAM3X code for edge turbulence fluid simulations of tokamak plasmas in versatile magnetic geometries. *J. Comp. Phys.*, 321:606–623, 2016.

[TGT+09] P. Tamain, Ph. Ghendrih, E. Tsitrone, Y. Sarazin, X. Garbet, V. Grandgirard, J. Gunn, E. Serre, G. Ciraolo, and G. Chiavassa. 3d modelling of edge parallel flow asymmetries. *J. of Nuc. Mat.*, 390-391:347–350, 2009.

[TGT+10] P. Tamain, Ph. Ghendrih, E. Tsitrone, V. Grandgirard, X. Garbet, Y. Sarazin, E. Serre, G. Ciraolo, and G. Chiavassa. Tokam-3d: A 3d fluid code for transport and turbulence in the edge plasma of tokamaks. *J. Comp. Phys.*, 229:361–378, 2010.

[TS01] Ulrich Trottenberg and Anton Schuller. *Multigrid*. Academic Press, Inc., Orlando, FL, USA, 2001.

[UDR05] M.V. Umansky, M.S. Day, and T.D. Rognlien. On numerical solution of strongly anisotropic diffusion equation on misaligned grids. *Num. Heat Transfer, B*, 47:533–554, 2005.

[vEKdB14] B. van Es, B. Koren, and H.J. de Blank. Finite-difference schemes for anisotropic diffusion. *J. Comp. Phys.*, 272:526–549, 2014.

[Wes92] P. Wesseling. An introduction to multigrid methods. *John Wiles & Sons*, 1992.

[Wes97] J. Wesson. Tokamaks. *Clarendon, Oxford Univ. Press*, 1997.

Appendix A

Sensitivity of the elliptic problem to μ in a periodic domain

The following elliptic boundary value problem has been considered in this work with $\mu = 1$:

$$-\boldsymbol{\nabla} \cdot \boldsymbol{\mathcal{K}} \boldsymbol{\nabla} T + \mu T = S \quad \text{in } \Omega = [0, 2\pi] \times [0, 2\pi],$$

A non-zero positive value of μ allows us to consider a periodic computational domain, and so to separate the study of the discretization of the solution at the boundary with the one in the interior of the domain. Here, we show that the results presented in the paper are little sensitive to the value of μ. Fig. A.1 shows indeed that the $\mathcal{H}^1$-error converges whatever the value of μ, for $\mu \in [10^{-6}, 1]$. Obviously, when μ reaches near zero values, the problem above tends to become singular (Poisson's equation) in the periodic domain, and the resolution of the problem becomes much more demanding, what explains the increasing number of grid points needed to converge when μ decreases.

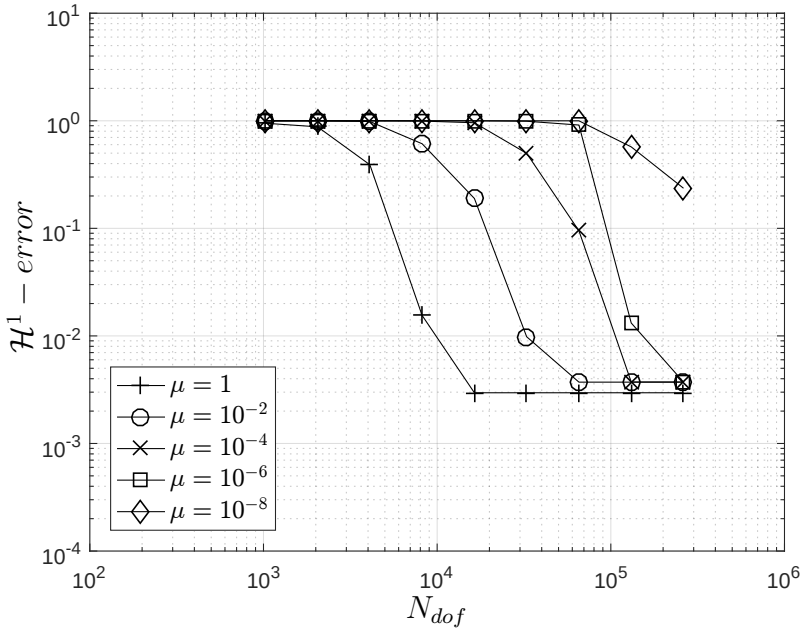


Figure A.1: $\mathcal{H}^1$-error with respect to the resolution N_{dof} for $\mu \in [10^{-6}, 1]$. Bi-periodic domain with $K_{b_\parallel} = 1$ and $K_{b_\perp} = 0$.

Appendix B

Impact of resolution on the representation of the solution

The aligned discretizations examined in the thesis are non-standard in the context finite differences, and we provide more information on the impact of spatial resolution on the representation of the solution of problems examined.

B.1 The Nyquist-Shannon theorem

For standard finite-differences, the solution is locally projected on a local polynomial basis where the two-dimensional functions correspond to the product of one-dimensional basis functions. In this case, the uniformity of the grid enables Fourier transform of the function considered, and it can be shown that the function will be adequately represented if the grid satisfies the one-dimensional Nyquist-Shannon theorem in each direction of the grid, based on the corresponding maximum wavenumber. In this context, the minimum number of degrees of freedom required to represent the function is then the product of the number of points required in each direction.

Aligned discretizations have the advantage of giving a better representation of an aligned function in two and three dimensions, with more adapted basis functions. The discussion of the required number of grid points is then more elaborate [PM62]. It is clear that these discretizations enable a reduction of the number of grid points to adequately represent aligned functions compared to two- or three-dimensional lattices using discretizations with standard (non aligned) finite differences. They, however, present us with a technical difficulty, which lies in the fact that "grid points" are to be seen as sampling points of the considered function: knowledge of the function itself, or its approximation, can only be obtained once the knowledge of sampling values has been augmented by the knowledge of the basis functions considered. Evaluation of the error by the classical discrete $\mathcal{L}^2$-norm is then insufficient since it will only evaluate the error

in the sampling values of the solution. As a result, it can completely fail to recognize aliasing errors. This difficulty is then circumvented by also evaluating the error in sampling values of the derivatives using the $\mathcal{H}^1$-norm defined in Appendix C, which can only be evaluated if the used basis functions are known.

In the following, we discuss the observed error in solutions of an anisotropic Helmholtz problem, where the solution is strongly aligned. Results are discussed here with respect to the Nyquist-Shannon theorem [Nyq28, Sha48], which provides the minimal resolution required to accurately represent the solution, i.e. $2m$ in each direction, where m is the highest wavenumber of the solution in this direction. The $\mathcal{L}^2$ and $\mathcal{H}^1$ norms of the error, respectively $\|T-T_a\|_{\mathcal{L}^2}$ and $\|T-T_a\|_{\mathcal{H}^1}$, are plotted on Fig. B.1 for all numerical schemes.

For *non-aligned methods*, Fig. B.1a shows as expected that below the Nyquist-Shannon resolution (dotted line, $N_y = 2m_y$), the resolution is not fine enough to accurately represent the solution. Aliasing effects may eventually lead to misleadingly small values of the $\mathcal{L}^2$ error in the solution of elliptic problems considered, observed here for $N_{dof} = 2.88\times10^2$ with the *Günter's method*. For larger resolutions ($N_y > 2m_y$) all the errors dominated by the discretization error in the y-direction decrease when increasing N_y. The minimal value is reached for a resolution corresponding to a perfect alignment of the grid with the solution (dashed line), i.e. for $N_y = N_x/\tan^{-1}(\alpha)$. In the present case, 3 grid points of the 9-point stencil used in the *Günter's method* are exactly aligned with $\mathbf{b}$, and the stencil for the parallel Laplacian actually reduces to three points along the main diffusion direction. The parallel Laplacian of the aligned fluctuations is thus exactly zero at the discrete level, and the aligned fluctuations are treated exactly. Beyond, the resolution in the x-direction being fixed, the discretization error in this direction becomes dominant and increases whatever the resolution used in the y-direction.

For *aligned methods* (Fig. B.1b), oriented stencils need an interpolation step in the y-direction to evaluate the parallel derivative introducing an additional discretization error related to the finite-difference scheme that can be large if the resolution is smaller than the Nyquist-Shannon resolution. For both resolutions in the x-direction, the error decreases when increasing the resolution in the y-direction. Oscillations of the error associated to local minima and maxima corresponding to resolutions for which the grids are aligned (minima), when the "diffusion line" going through one grid point also intercepts another grid point, or the most misaligned (maxima) along the parallel diffusion direction. When using a finite-difference discretization, the interpolation error being proportional to $1/d_\parallel^2 = 1/(\Delta x \cos^{-1}\alpha)^2$, where $d_\parallel$ is defined in Eq. 5.33, it increases when N_x increases (i.e. $d_\parallel$ decreases) for the same number of grid points in the y-direction as shown on Fig. B.1b.

Figure B.1a in particular illustrates the difficulties associated with the use of the discrete $\mathcal{L}^2$-norm and justifies the use of the $\mathcal{H}^1$-error in the analysis of the accuracy tests.

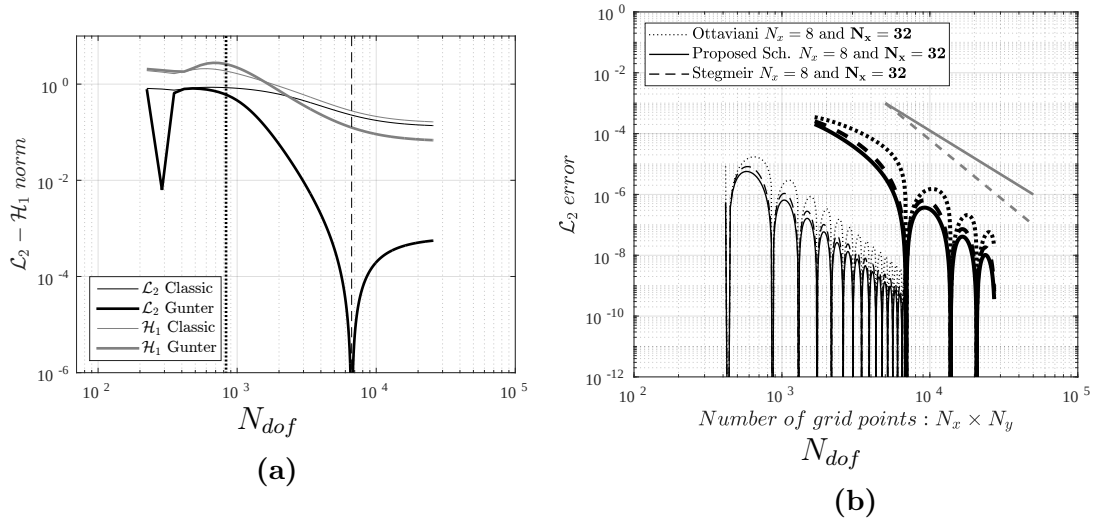


Figure B.1: $\|T - T_a\|_{\mathcal{L}^2}$ and $\|T - T_a\|_{\mathcal{H}^1}$ errors when increasing resolution in the y-direction, $N_y \in [50, 250]$. (a) *Non-aligned methods*, $N_x = 32$. The dotted line corresponds to the minimal resolution prescribed by the Nyquist-Shannon theorem. The dashed line corresponds to the resolution for which the grid is perfectly aligned with the direction of the parallel diffusion. (b) *Aligned methods*, $N_x = 8$ (thin lines) and $N_x = 32$ (thick lines). T_a (Eq. 3.45) is defined with $C_1 = C_3 = 0$, $C_2 = 1$ and with $m_y = 13$ and $m_{x,1} = 2$ leading to $\alpha = 8.75°$. With these values, the field remains constant along the parallel direction defined by $\mathbf{b} = (\cos\alpha, \sin\alpha, 0)$, while rapid variations can be observed in the perpendicular direction.

B.2 Modal analysis

A modal representation of highly anisotropic flows allow us to deepen the previous analysis on the signal treatment given by aligned methods. Fig. B.2a shows a representative anisotropic flow field, and its pseudo-spectrum (FFT, Fig. B.2b) for a 128×128 grid. The modes for a given $T(x, y)$ test case can be defined by the expression:

$$T(x, y) = \sum_{mn} T_{mn} e^{i(mx+ny)} \tag{B.1}$$

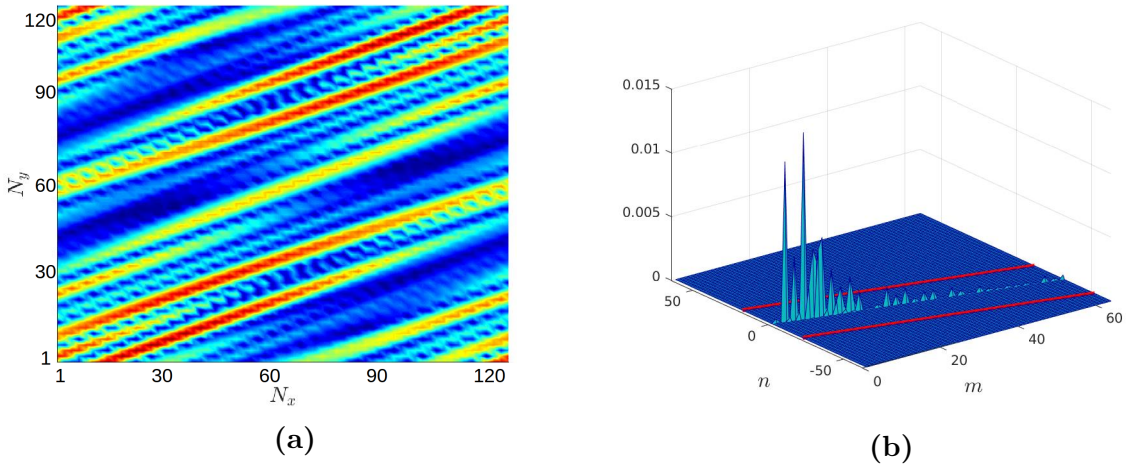
with $m = -N_x/2, ..., -1, 0, 1, ..., N_x/2$ and $n = -N_y/2, ..., -1, 0, 1, ..., N_y/2$ the modes represented in the grid. The T_{mn} modes show a strong orientation conditioned by the diffusion direction: the modes non aligned with $\mathbf{b}$ are rapidly damped. So one can establish that Fig. B.2 is representative of developed highly anisotropic flows.

The non aligned schemes introduced in Sec. 2 are used to discretize Eq. B.1 for any resolution in x or y direction. However, Fig. B.2 shows the spectral energy of the solution $T(x, y)$ in anisotropic diffusion problems. One can identify aligned, or nearly aligned fluctuations are carrying most of the energy content. In this representation, it is clear that the relevant modes are by far fewer than those described on the grid, opening

the possibility to reduce the number of unknowns by focusing on nearly aligned modes. One can for example use the alternative expression of Eq. B.1:

$$T(x,y) = \sum_{m'=-\frac{N_x}{2}}^{\frac{N_x}{2}} \sum_{n=-\frac{N_y}{2}}^{\frac{N_y}{2}} T_{m'n} e^{inx} e^{im'(y-x\tan(\alpha))} \tag{B.2}$$

where m' are the modes in the y-direction. Here, $N_y/2\tan(\alpha) \pm N_x + 2$ denotes the band width representing the aligned modes (red lines in Fig. B.2b in $N_x = 16$), and the parallel variations are captured by variations in x. The slow, or vanishing, parallel variations of solutions of strongly anisotropic problems considered then opens to the possibility of reducing N_x without loss of quality of the representation. Actually, the aligned schemes presented here discretize the modes described by Eq. B.2. A full detailed analysis is presented by Ottaviani in [Ott11].



(a) **(b)**

Figure B.2: (a) 2D plot of the field fluctuations in the physical space. (b) Energy of the modes in the spectral space. Most of the total energy of the system is located in $m = n/tan(\alpha)$, that opens to the possibility to reduce the grid in the x-direction, represented here by the minimal Nyquist-Shannon criterion ($\pm N_y/(2\tan(\alpha))$ plus a certainty factor $\pm N_x/2$ (red lines).

B.3 Eigenvalues and eigenvectors in anisotropic problems

The Helmholtz operator $\mathcal{A} = \mu - \mathcal{D}_{b\|}$ in bi-periodic domain is bi-periodic and uniform. It inherits properties of $\mathcal{D}_{b\|}$ and is self-adjoint. Likewise, the discretizations of $\mathcal{A}$, matrices $A_{Classic}$, A_{Gunter}, $A_{Ottaviani}$, $A_{Stegmeir}$ and $A_{Present\ S.}$, are bi-periodic and symmetric. The

uniformity of the operator, and bi-periodicity of the solution results in the bi-periodicity of all matrices discretizing $\mathcal{A}$, which in turn leads to the property that the eigenvectors of all matrices discretizing $\mathcal{A}$ considered are discrete, two-dimensional Fourier modes. Given a $N_x \times N_y$ grid, and for $j \in \{1, ..., N_x\}$ and $k \in \{1, ..., N_y\}$, the eigenvectors are given by:

$$e_{ij}^{mn} = \exp\left\{ i \left[m\frac{(j-1)2\pi}{N_x} + n\frac{(k-1)2\pi}{N_y} \right] \right\} \tag{B.3}$$

with $i = \sqrt{-1}$, $m \in \{-N_x/2 + 1, ..., 0, ..., N_x/2\}$ and $n \in \{-N_y/2 + 1, ..., 0, ..., N_y/2\}$. The eigenvalues λ being solution of:

$$A\ e_{ij}^{mn} = \lambda\ e_{ij}^{mn} \tag{B.4}$$

one can straightforwardly obtain the $(N_x N_y)/2$ eigenvalues: the eigenvectors of any matrix discretizing $\mathcal{A}$ on the uniform, bi-periodic grid are Fourier modes and the eigenvalues can be obtained by:

$$|\lambda^{mn}| = \frac{\|v_{ij}^{mn}\|}{\|e_{ij}^{mn}\|};\ where\ v_{ij}^{mn} = A\ e_{ij}^{mn} \tag{B.5}$$

The eigenvectors are identical for all discretizations and one can obtain:

- for all m, n: the exact values of λ by standard Fourier analysis,
- for all m, n: the modulus of λ of $A_{Classic}$, A_{Gunter}, $A_{Ottaviani}$, $A_{Stegmeir}$ and $A_{Present\ S.}$ using Eqs. B.3, B.4 and B.5.

Fig. B.3 represents the numerical values of λ with respect to the theoretical ones for non aligned (Fig. B.3a) and aligned methods (Fig. B.3b) in a reduced N_x grid. Results show how non aligned methods fail in the representation of the values of λ corresponding to aligned modes. This is to be expected since these modes are aliased owing to the under-sampling of variations in the x-direction by the grid (represented in Fig. B.3a by the gray slashed line). Contrary to the non aligned methods, aligned methods achieve a much better approximation of the theoretical values of λ for grids with $N_y > N_x$ (gray line).

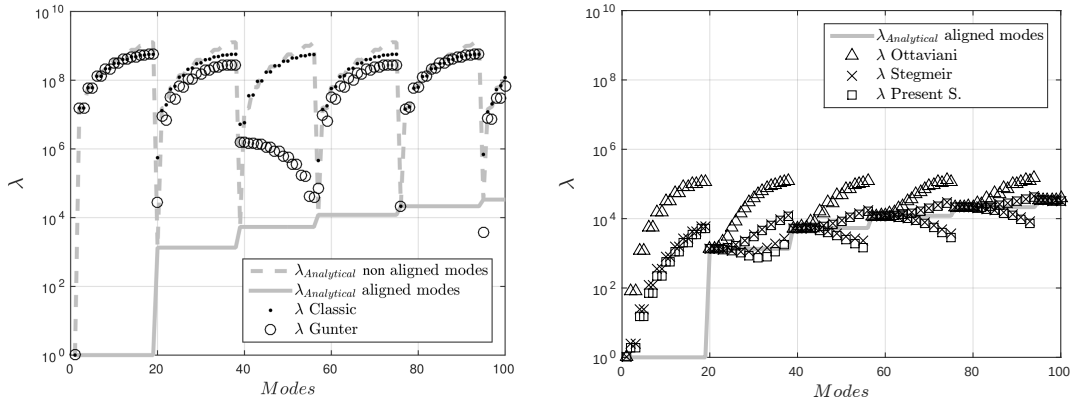


Figure B.3: Eigenvalues obtained with non aligned approaches (a), and aligned approaches (b) for $K_{b\|} = 10^6$ and a grid with $N_x = 4$ a,d $N_y = 512$. The gray line represents the theoretical eigenvalues of modes oriented with the pitch angle α (oriented structures of the solution). The gray dashed line shows the alised modes in the grid nodes.

Appendix C

Discretization of the gradient in the $\mathcal{H}^1$-error

The $\mathcal{H}^1$-error is used in the paper to avoid the shortcomings of the discrete $\mathcal{L}^2$-norm (Appendix B). The $\mathcal{H}^1$-error defines as:

$$\begin{aligned} \|T - T_a\|^2_{\mathcal{H}^1} &= \|T - T_a\|^2_{\mathcal{L}^2} + \|\nabla(T - T_a)\|^2_{\mathcal{L}^2} \\ &\geqslant \|T - T_a\|^2_{\mathcal{L}^2} + \|\nabla_x(T - T_a)\|^2_{\mathcal{L}^2} + \|\nabla_y(T - T_a)\|^2_{\mathcal{L}^2} \end{aligned}$$

It requires the evaluation of the error in the gradients in each direction x and y. Depending on the method used for the discretization, different stencils are used:

- For the *classical method*, gradients are evaluated by finite differences from grid points located in both x and y directions, Fig. C.1a. The gradients simply express as:

$$\nabla_x T_{ij} \approx \frac{T_{i+1j} - T_{i-1j}}{2\Delta x}, \quad \nabla_y T_{ij} \approx \frac{T_{ij+1} - T_{ij-1}}{2\Delta y}, \tag{C.1}$$

- For the *Günter's method*, the stencils involve the values of the function at the center of the surrounding cells Fig. C.1b such as:

$$\begin{aligned} \nabla_x T_{i\,j} &\approx \frac{1}{2}\left(\frac{T^{int}_{i+\frac{1}{2}\,j+\frac{1}{2}} + T^{int}_{i+\frac{1}{2}\,j-\frac{1}{2}}}{\Delta x} - \frac{T_{i-\frac{1}{2}\,j+\frac{1}{2}} + T^{int}_{i-\frac{1}{2}\,j-\frac{1}{2}}}{\Delta x}\right) \\ \nabla_y T_{i\,j} &\approx \frac{1}{2}\left(\frac{T^{int}_{i+\frac{1}{2}\,j+\frac{1}{2}} + T^{int}_{i-\frac{1}{2}\,j+\frac{1}{2}}}{\Delta y} - \frac{T^{int}_{i+\frac{1}{2}\,j-\frac{1}{2}} + T^{int}_{i-\frac{1}{2}\,j-\frac{1}{2}}}{\Delta y}\right), \end{aligned} \tag{C.2}$$

  where T^{int} are evaluated from the nearest grid points as follows:

$$T^{int}_{i+\frac{1}{2}\,j+\frac{1}{2}} = \frac{T^{int}_{i+1\,j+1} + T^{int}_{i\,j+1} + T^{int}_{i+1\,j} + T^{int}_{i\,j}}{4} \tag{C.3}$$

- For the *aligned methods*, gradients in the x and y directions are obtained from the gradients evaluated in the parallel and perpendicular directions as detailed in Chapter 3. Thus:

$$\nabla_{\parallel} T_{ij} \approx \frac{T^{int+} - T^{int-}}{d_{\parallel}|_{i-1}^{i} + d_{\parallel}|_{i}^{i+1}} \qquad \nabla_{\perp} T_{ij} \approx \frac{\nabla_y T_{ij} - \nabla_{\parallel} T_{ij} \sin\alpha}{\cos\alpha}, \tag{C.4}$$

and so (Fig. 3.9):

$$\nabla_x T_{ij} \approx \nabla_{\parallel} T \cos(\alpha) - \nabla_{\perp} T \sin(\alpha), \qquad \nabla_y T_{i,j} \approx \nabla_{\parallel} T \sin(\alpha) + \nabla_{\perp} T \cos(\alpha), \tag{C.5}$$

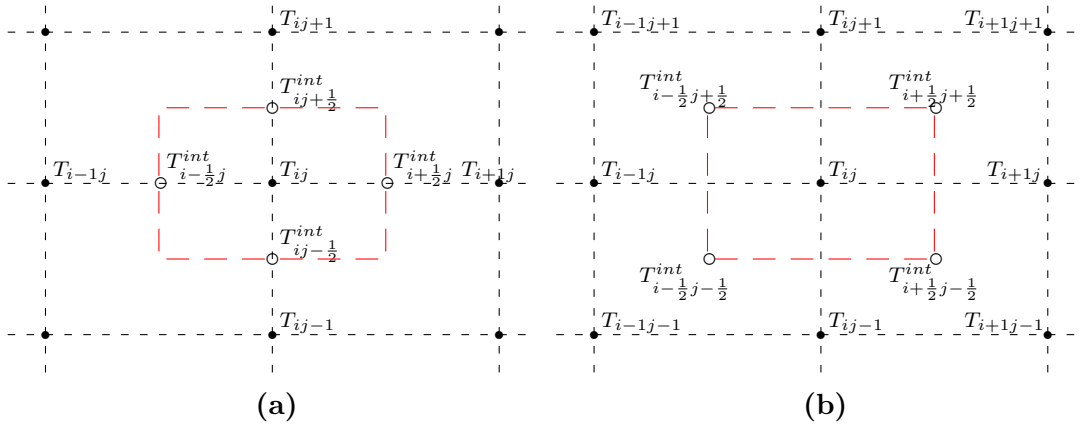


Figure C.1: Examples of stencils. (a) The *classical method*. (b) *Günter's method*.

Appendix D

Details on solving the linear system

All the results presented in this paper require the inversion of a discrete matrix. The solution of the linear system Eq. 2.1 is here calculated by the Matlab's *backslash* function for unsymmetric sparse linear systems (UMFPACK [Dav04]).

The high anisotropy can however lead to a ill-conditioned discrete matrix, since the diagonal scales as $K_{b\|}$. $K_{b\|}$ being imposed by the physics, the condition number is a function of the number of degrees of freedom and of discretization scheme used for the Laplacian.

We analyse here the value of condition number depending on the scheme used for solving Eq. 2.1 in a bi-periodic 2D domain. In Figs. D.1, we compare for both the Günter's (non-aligned) (a) and the present (aligned) (b) methods the values of the condition number in function of the distribution of the points between the x and y-directions, keeping constant $N_{d.o.f.}$. The values of the condition number with the *present method* are several orders below the values obtained with the *Günter's method*. Moreover, the results show that the distribution of points impacts the value of the condition number, but differently depending on the method. For the *Günter's method*, the values decrease with N_x increasing to 64, the minimum value being obtained for a 64×512 grid. For the *present method*, it is the opposite, the minimum value being reached for a 8×4096 grid.

In Fig. D.2a, the values of the condition number are shown when increasing the $K_{b\|}$, considering a 64×512 grid for all non-aligned methods, and a 8×4096 grid for all the aligned methods. The condition number grows linearly with $K_{b\|}$ for all methods. However, its values are two orders lower when using aligned methods. This is an important feature when studying extreme values of $K_{b\|}$. The grid distribution has no impact here on the results for all the aligned methods.

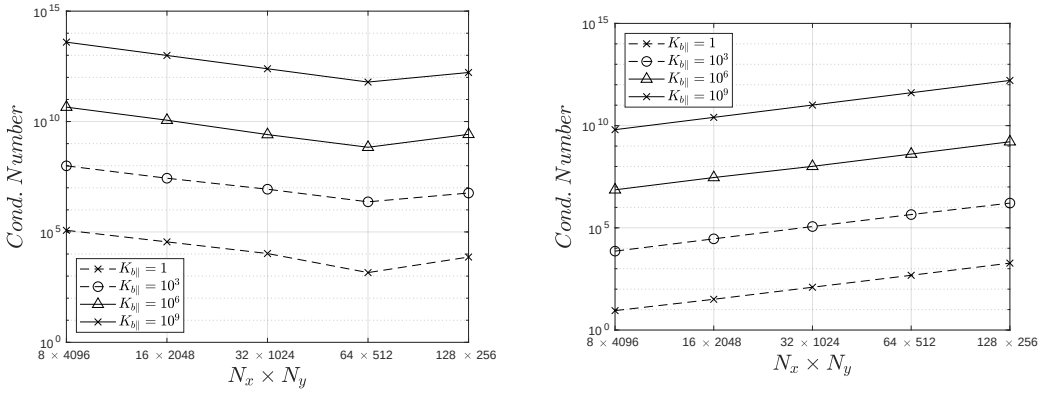


Figure D.1: Evolution of the values of the condition number for different grid points distribution and for different values of the parallel diffusion. N_x and N_y are varied keeping constant $N_{d.o.f.}$. The *Günter's method* (a) and the *present method* (b).

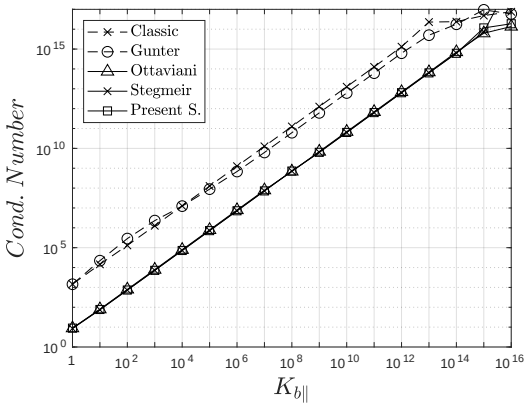


Figure D.2: Values of condition number for all methods when increasing $K_{b\|}$. The grids are 64×512 and 8×4096 for non-aligned and aligned methods, respectively.

Appendix E

Résumé de thèse

E.1 Introduction et motivation

Les problèmes elliptiques hautement anisotropes se présentent dans de nombreux modèles physiques qui doivent être résolus numériquement. Une direction de diffusion dominante est alors introduite (appelée ici direction parallèle) le long de laquelle le coefficient de diffusion est plusieurs ordres de grandeur plus grands que dans la direction perpendiculaire. Dans ce cas, les méthodes aux différences finies standard ne sont pas conçues pour fournir une discrétisation optimale et peuvent conduire à une diffusion perpendiculaire artificielle potentiellement importante, résultant d'erreurs dans l'approximation de la diffusion parallèle.

Cette thèse se concentre sur trois axes principaux pour résoudre les équations elliptiques anisotropes de manière appropriée : un schéma aligné et conservatif de différences finies pour discrétiser l'opérateur Laplacien, une reformulation de l'équation de Helmholtz pour réduire la diffusion numérique, et un solveur basé sur les méthodes multi-grille comme préconditionneur d'un solveur GMRES. Les deux premiers chapitres sont consacrés à la présentation du cadre de cette thèse.

Au chapitre 1, une brève introduction à la fusion par confinement magnétique est présentée, identifiant les problèmes numériques soulevés par la résolution des équations fluides, en particulier dans la région proche au bord (Scrape-Off-Layer). Le problème numérique que nous allons traiter est essentiellement un problème elliptique anisotrope où la diffusion est de 5 à 8 ordres de grandeur plus grande dans la direction parallèle que dans la direction perpendiculaire.

Dans le chapitre 2, une introduction bibliographique aux méthodes numériques résolvant les équations elliptiques anisotropes est présentée, avec un accent sur les méthodes aux différences finies.

Dans le chapitre 3, un schéma de discrétisation aligné est proposé en utilisant des grilles cartésiennes non alignées. Selon la méthode Support Operator Method (SOM),

la propriété d'auto-ajoint de l'opérateur de diffusion parallèle est maintenue au niveau discret. Par rapport aux méthodes existantes, la formulation actuelle garantit la conservation des flux dans des directions parallèles et perpendiculaires. De plus, dans les domaines bornés, une discrétisation des conditions aux limites est présentée afin d'assurer une précision comparable de la solution. Des tests numériques basés sur des solutions manufacturées montrent que la méthode est capable de fournir des approximations numériques précises et stables dans des domaines périodiques et bornés avec un nombre considérablement réduit de degrés de liberté par rapport autres approches non aligneés.

Une reformulation de l'équation de Helmholtz est présentée au chapitre 4 pour limiter la diffusion numérique liée à la discrétisation du Laplacien pour les valeurs élevées de diffusion parallèle. La méthode est basée sur la séparation de la solution en une partie alignée et non alignée, par rapport à l'opérateur de diffusion parallèle, grâce à des méthodes de filtrage. Les cas de tests montrent que cette reformulation de l'équation de Helmholtz élimine la diffusion perpendiculaire numérique, avec une efficacité d'autant plus accrue que les valeurs de diffusivité parallèle sont élevées.

Afin de résoudre efficacement les équations anisotropes elliptiques pour les grands systèmes d'équations, un solveur itératif basé sur des algorithmes multi-grilles géométriques est proposé au chapitre 5. Cet algorithme est plus tard posé comme préconditionneur d'un solveur GMRES, exhibant une réduction drastique du temps et de la mémoire requise par rapport à des solveurs directs résolvant les équations Helmholtz et Poisson, et ce pour différents types de conditions aux limites.

La thèse est conclue par une analyse critique des aspects numériques des discrétisations alignées étudiées. Une attention particulière est accordée à l'application des méthodes étudiées dans les codes de turbulence plasma 3D, tels que TOKAM3X développé par le CEA.

E.2 Limitations numériques en la discrétisation du code fluide TOKAM3X

E.2.1 Introduction

Ce travail concerne le développement de schémas numériques avancés pour la simulation de plasmas pour la fusion magnétique. Ce chapitre décrit brièvement les principaux principes physiques de la fusion, et fournit l'ensemble des équations de fluides implémentées dans TOKAM3X [TBC+16], le code actuellement développé par l'équipe pour simuler les flux de plasma au bord de la chambre du tokamak. Une des spécificités de cette configuration est la forte anisotropie introduite par le champ magnétique entre sa direction parallèle et transversale dans le flux de plasma. Mathématiquement, cela conduit à fortement anisotrope opérateurs différentiels dans les équations. Ce travail est

motivé par le besoin de concevoir des schémas numériques plus efficaces permettant de satisfaire les exigences de résolution et de précision pour effectuer des simulations fiables de turbulence dans des configurations magnétiques réalistes. Le lecteur est référé aux ouvrages de la Réf. [Wes97, Fre07, GR07, Dav01] pour plus d'informations.

E.2.2 Le modèle fluide de TOKAM3X

Une bonne compréhension du bord nécessiterait *complète-f gyrokinetiques* simulations basées sur la fonction de distribution. Des simulations gyrokinetiques complètes pionnières du début de pointe apparaissent dans la communauté de la fusion, abordant les phénomènes physiques d'intérêt fondamental pour le fonctionnement de la fusion comme la formation de barrières de transport [CKG$^+$06, CKT$^+$17]. Cependant, malgré la croissance exponentielle de la vitesse des ordinateurs et les améliorations importantes apportées à la technologie informatique, ils demeurent extrêmement coûteux du point de vue informatique. C'est particulièrement vrai dans la région proche de la paroi, où la recirculation des particules nécessite d'aborder la dynamique des électrons et des ions sur le même pied, et dans une topologie magnétique beaucoup plus complexe que dans le noyau. Par conséquent, bien qu'approximative, l'approche par fluide demeure une approche standard près du mur où la température est plus basse et le parcours libre moyenne radiative est beaucoup plus petite que dans le noyau.

Des équations tridimensionnelles de conservation des fluides sont obtenues pour les électrons et les ions en utilisant des fermetures simplifiées développées par Braginskii [Bra65]. Le modèle présenté ci-dessous est celui mis en œuvre dans la version isotherme du code TOKAM3X développé depuis de nombreuses années par le laboratoire en étroite collaboration avec le CEA, voir dans la référence [Tam07, TBC$^+$16, Col15, GTB$^+$17, TGT$^+$10]. Sous certaines hypothèses et ordre détaillés ci-dessous, quatre équations sont dérivées pour quatre champs sans dimension inconnus : la densité électronique N, l'élan parallèle ionique Γ, le potentiel électrostatique Φ et le courant parallèle $j_{b\parallel}$ qui définit la vitesse d'advection parallèle des électrons.

E.2.3 Equations fluides

Les hypothèses et l'ordre énoncés dans Sec. 1.2.1 mènent à une équation de conservation pour la densité électronique N (par souci de simplicité par rapport à l'équation de densité ionique), pour la dynamique parallèle ionique Γ (obtenu par la somme des équations pour les ions et les électrons) et pour la vorticité W, qui remplace l'équation de la balance des charges ($\nabla \mathbf{j} = 0$), voir dans Ref. [TBC$^+$16] :

$$
\begin{cases}
\partial_t N + \nabla \cdot \left(N \mathbf{u}^e \right) = S_N + \nabla \cdot \left(D_N \nabla_{b\perp} N \right) \\
\partial_t \Gamma + \nabla \cdot \left(\Gamma \mathbf{u}^i \right) = -\nabla_{b\|} P + \nabla \cdot \left(D_\Gamma \nabla_{b\perp} \Gamma \right) \\
\partial_t W + \nabla \cdot \left(W \mathbf{u}^i \right) = \nabla \cdot \left(N(\mathbf{u}^i_{\nabla B} - \mathbf{u}^e_{\nabla B}) + j_{b\|} \mathbf{b} \right) + \nabla \cdot \left(D_W \nabla_{b\perp} W \right) \\
with \\
j_{b\|} = -\frac{1}{\eta_{b\|}} \nabla_{b\|} \phi + \frac{1}{N \eta_{b\|}} \nabla_{b\|} N \\
W = \nabla \cdot \left(\frac{1}{B^2} \left(\nabla_{b\perp} \phi + \frac{1}{N} \nabla_{b\perp} N \right) \right)
\end{cases}
\tag{E.1}
$$

où $\eta_\|$ est la résistivité radiative parallèle normalisée du plasma, S_N un terme source volumétrique inclus pour entraîner le flux de particules, et $\nabla_{b\|}$ et $\nabla_{b\perp}$ sont les gradients parallèles et perpendiculaires respectivement à $\mathbf{b}$. Le terme d'entraînement dans l'équation de momentum est le gradient de pression statique parallèle sans dimension $\nabla_{b\|} P$.

Dans toutes les équations ci-dessus, les termes de diffusion efficaces tiennent en compte le transport radiatif et modélisent grosso modo l'effet des petites échelles turbulentes (plus petites que l'espacement entre les grilles) dans la direction du champ transversal, voir fig. E.1. Leur composant parallèle a été négligé par rapport à la convection parallèle. $D_{N,\Gamma,W}$ sont des constantes arbitraires, généralement inférieures à 1.

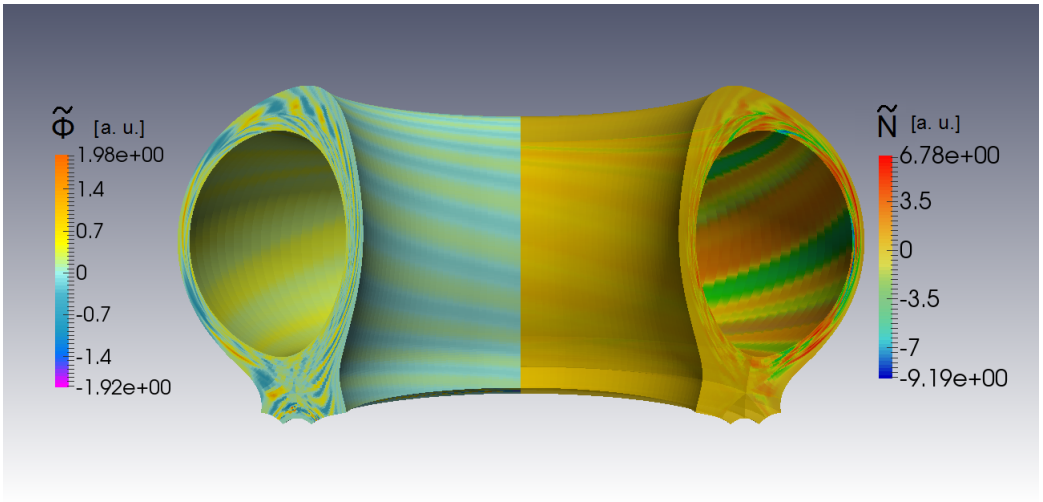


Figure E.1: Un exemple de structures turbulentes dans une simulation TOKAM3X dans une configuration de déviation [TBC+16, TGT+09]. Le transport et la diffusion hautement anisotropiques génèrent un flux caractérisé par des structures allongées le long des lignes de champ magnétique, en combinaison avec des variations spatiales rapides dans la direction perpendiculaire dues à la turbulence. Fluctuations du potentiel électrique (à gauche) et de la densité (à droite).

E.2.4 Schémas numériques

Comme mentionné ci-dessus, le flux dans le sens parallèle correspond à un flux de gaz compressible, alors que dans le sens perpendiculaire il correspond à un flux quasi incompressible en raison du fort champ magnétique, dominé par des processus turbulents. En outre, la topologie magnétique du bord du tokamak est complexe et rend le flux fortement anisotrope. Par conséquent, les équations de conservation présentées ci-dessus et gouvernant le plasma de bord/SOL nécessitent des algorithmes spécifiques qui habituellement divisent la discrétisation des directions parallèles et perpendiculaires.

Afin de limiter la diffusion numérique, les équations ci-dessus sont discrétisées sur une grille structurée de surface de flux magnétique alignée où la première direction du maillage est le long de l'iso-ψ lignes dans le plan poloidal (voir 2 exemples dans Figs. E.2 et E.3). Ici, ψ indique la coordonnée orthogonale des surfaces magnétiques, qui correspond à la coordonnée radiale r en coupe circulaire, voir [TGT+09].

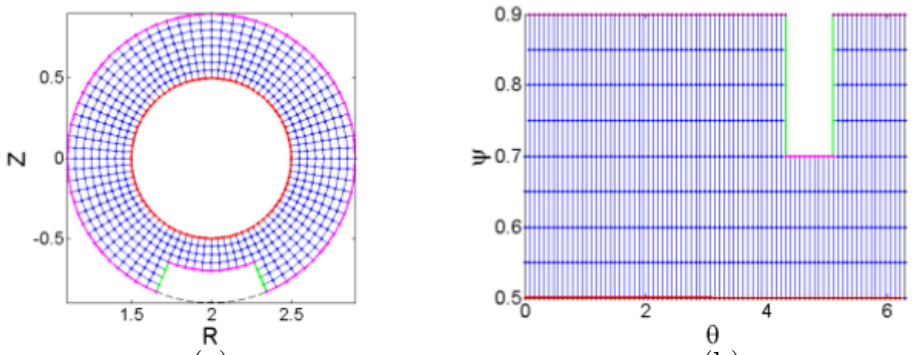


Figure E.2: Exemples de mailles en section circulaire limitée. (a) Répartition des mailles dans le physique (R, Z)-plane (à gauche) et dans le (ψ, θ)-plane (à droite). Le limiteur est situé au bas de la machine.

La discrétisation spatiale est basée sur un schéma conservatif de deuxième ordre de différences finies associé à une reconstruction WENO de troisième ordre pour les termes d'advection, pour traiter à la fois les chocs et les structures compliquées de la solution [LOC94]. L'évolution temporelle est basée sur un fractionnement d'opérateur de premier ordre. Les termes implicites et explicites sont les suivants :

- Les termes advection et source sont principalement non-linéaires. Leur dynamique est sur une échelle de temps ionique qui permet un avancement explicite.
- Les termes de courant parallèle expriment l'évolution du potentiel électrique du plasma. Ils sont avancés en utilisant un solveur 3D entièrement implicite afin

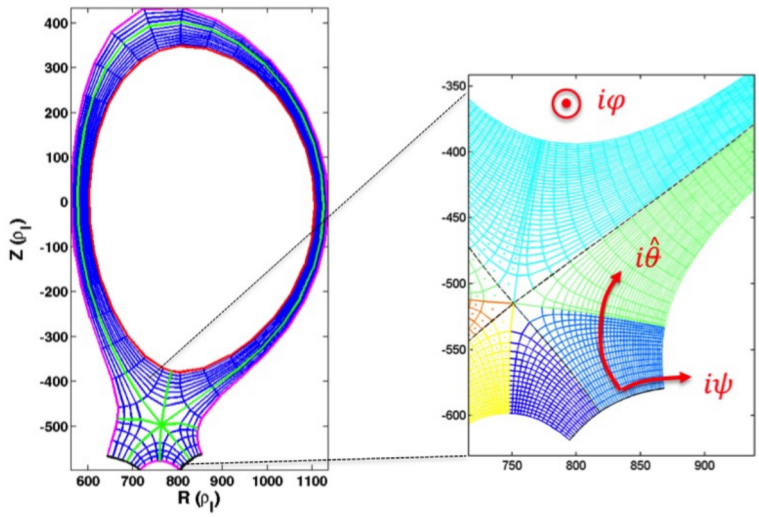


Figure E.3: Exemple de maille en coupe poloïdal-radial avec diverteur. Répartition du maillage dans l'espace physique d'une géométrie avec diverteur mettant l'accent sur la décomposition du domaine (lignes vertes) et la répartition de la grille autour du point X.

de capturer la dynamique rapide associée sans limiter considérablement le pas de temps.

- Les termes de diffusion perpendiculaire sont avancés implicitement afin de permettre un grand coefficient de diffusion, en exécutant le code en mode transport (i.e. pas de petites échelles turbulentes).

Pour tous ces termes, l'avancement des termes courants parallèles dans l'équation de vorticité est la principale question numérique. L'évolution temporelle du potentiel plasmatique s'écrit comme suit :

$$(\mathcal{L}^{b\perp} + \ \delta t\ \mathcal{L}^{b\|})\phi^{\star\star} = W^{\star} - \mathcal{L}^{b\perp}N^{\star} + \ \delta t\ \mathcal{L}^{b\|} lnN^{\star} \tag{E.2}$$

où $\mathcal{L}^{b\perp,b\|}$ sont des opérateurs différentiels spatiaux définis comme $\mathcal{L}^{b\perp} = \nabla\cdot(\frac{1}{B^2}\nabla_{b\perp}\cdot)$ et $\mathcal{L}^{b\|} = \nabla\cdot(\frac{1}{\eta_{b\|}}b\nabla_{b\|}\cdot)$

L'eq. E.2 associé à l'ensemble des conditions limites complexes détaillées ci-dessus, illustre parfaitement le type de problème numérique que nous voulons aborder dans ce travail de doctorat. La très faible valeur de la résistivité parallèle dans le plasma tokamak ($\eta_{b\|} \approx 10^{-5} - 10^{-8}$ (valeurs normalisées)) conduit à un opérateur différentiel fortement anisotrope et très mal conditionné. La discrétisation d'un tel opérateur nécessite un schéma numérique bien adapté pour limiter la diffusion numérique sans faire un effort inabordable sur le nombre de degré de liberté qui rendra le calcul trop coûteux.

Cette interaction plasma-paroi implique également des processus atomiques complexes qui conduisent à des problèmes multi-physiques et donc à des questions numériques difficiles pour modéliser correctement le flux de plasma dans cette région.

E.3 Schéma de differences finites conservatif pour problèmes elliptiques anisotrophes en domaines bornés

E.3.1 Introduction

Les systèmes elliptiques en derivés partiels sont omniprésents dans les modèles physiques et les simulations numériques. On les retrouve dans les modèles de fluides utilisés en mécanique, en géophysique, en physique des plasmas, mais aussi dans d'autres domaines de recherche comme la microélectronique, l'optique, le traitement d'images, etc., la liste n'étant pas exhaustive.

Un problème typique à résoudre avec des conditions limites appropriées est l'équation de Poisson:

$$-\boldsymbol{\nabla} \cdot (\boldsymbol{\mathcal{K}} \cdot \boldsymbol{\nabla}) T = S, \quad \in \ \Omega \subset \mathbb{R}^3$$

où $\boldsymbol{\mathcal{K}}$ est le tenseur de diffusion. Dans de nombreuses configurations l'isotropie du problème peut être brisée, et une direction préférée (direction de l'anisotropie) est ainsi introduite, qui conduit à un tenseur de diffusion anisotrope. Puisque ce travail est motivé par des simulations de plasma de fusion [TBC+16, TGT+09], la direction de l'anisotropie est liée à l'anisotropie des composants de champ magnétique dans le tokamak et il est désigné comme la direction parallèle, en référence à la direction le long des lignes de champ magnétique. La direction de l'anisotropie est ainsi soutenue par un champ vectoriel sans divergence qui ne disparaît jamais. En supposant en outre que la diffusion est isotrope dans les directions perpendiculaires, conduit à l'expression suivante du tenseur dans un système de coordonnées dont les axes coïncident avec les directions principales :

$$\begin{bmatrix} K_{b_\parallel} & 0 & 0 \\ 0 & K_{b_\perp} & 0 \\ 0 & 0 & K_{b_\perp} \end{bmatrix}$$

where $K_{b_\parallel}$ and $K_{b_\perp}$ are functions of space, with $K_{b_\parallel}/K_{b_\perp} \gg 1$.

Les méthodes de maillage régulières et structurées pour les lois de conservation ne sont généralement pas conçues pour discréditer ces opérateurs anisotropes de manière optimale. Ils favorisent intrinsèquement des directions alignées avec des points de maille, ce qui peut introduire des erreurs systématiques, même à haute résolution, lorsque les directions principales du tenseur de diffusion ne sont pas alignées avec les axes de la grille.

La discrétisation peut en particulier produire une diffusion numérique fausse significative dans la direction orthogonale vers l'anisotropie, qui peut avoir un impact significatif sur la dynamique perpendiculaire [UDR05]. Comme il est rappelé dans van Es *et al.* [vEKdB14], d'autres problèmes peuvent également se poser car la non-positivité près des gradients élevés [SH07], stagnation ou même perte de la convergence de l'erreur de discrétisation avec le raffinement du maillage, voir [BS92, GYKL05]. De tels effets fallacieux peuvent être réduits en utilisant un schéma de haut degrée pour discréditer les opérateurs dans le sens de l'anisotropie, mais de plus grands stencils sont alors nécessaires.

E.3.2 Modèle mathématique

Nous nous concentrons sur la résolution de problèmes de diffusion fortement anisotropiques en utilisant la discrétisation de temps implicite de premier ordre , ou les équations de Poisson anisotropiques survenant lors de l'étude de solutions stationnaires des mêmes problèmes de diffusion. Ces deux problèmes peuvent être décrits par le problème suivant, générique, valeur limite elliptique :

$$\begin{cases} -\boldsymbol{\nabla} \cdot \boldsymbol{\mathcal{K}} \boldsymbol{\nabla} T + \mu T = S & \text{in } \Omega, \\ \beta \nabla_{b\|} T + \gamma T = g & \text{on } \Gamma, \end{cases} \tag{E.3}$$

où Ω est un domaine limité en $\mathbb{R}^3$ avec limite Γ, fourni avec une base orthonormale $(\mathbf{e}_x, \mathbf{e}_y, \mathbf{e}_z)$ associé aux coordonnées cartésiennes (x, y, z). Nous supposons que les variables du problème satisfont aux hypothèses habituelles d'ellipticité et de régularité. μ est une constante positive (ou nulle) et S est un terme source donné. Les coefficients *beta* et γ, et g sont utilisés pour définir les conditions limites générales, qui peuvent être de type Dirichlet ($\beta = 0$), Neumann ($\gamma = 0$) ou Robin ($\beta \neq 0, \gamma \neq 0$). Avec un tel problème, les domaines périodiques et délimités peuvent être considérés qui nous permet de déconnecter la discrétisation de l'équation à l'intérieur du domaine et aux frontières.

L'anisotropie du problème est prise en compte via la définition du tenseur de diffusion symétrique $\boldsymbol{\mathcal{K}}$, où le premier envalue $K_{b\|}$ est supposé la direction de diffusion dominante ($K_{b\|} \gg K_{b\perp}$), que nous pouvons identifier avec la valeur propre normalisé $\mathbf{b}$. Selon ces hypothèses, le système se lit comme suit :

$$-\boldsymbol{\nabla} \cdot \left[\boldsymbol{\mathcal{R}} \begin{bmatrix} K_{b_\|} & 0 & 0 \\ 0 & K_{b_\perp} & 0 \\ 0 & 0 & K_{b_\perp} \end{bmatrix} \boldsymbol{\mathcal{R}}^{-1} \right] \begin{Bmatrix} \partial T/\partial x \\ \partial T/\partial y \\ \partial T/\partial z \end{Bmatrix} + \mu T = S. \tag{E.4}$$

où $\boldsymbol{\mathcal{R}}$ définit une matrice de rotation 3D.

$\boldsymbol{b}$ peut être fonction de l'espace, et il est supposé ici être sans divergence. Notons que dans les problèmes où $\boldsymbol{b}$ ne serait pas sans divergence. Les gradients le long des directions parallèles et perpendiculaires sont alors définis comme $\nabla_{b\|} = \boldsymbol{b} \cdot \boldsymbol{\nabla}$ et $\boldsymbol{\nabla}_{b\perp} = \boldsymbol{\nabla} - \boldsymbol{b}\nabla_{b\|}$, respectivement.

De grandes simplifications peuvent être obtenues en définissant une base orthonormale constituée par les valeurs propres normalisés de $\boldsymbol{\mathcal{K}}$, à savoir $(\mathbf{b}, \mathbf{e}_\perp^1, textbfe_\perp^2)$, et le système de coordonnées alignées associé $(b_\parallel, b_\perp^1, b_\perp^2)$. L'orthogonalité des valeurs propres de $\boldsymbol{\mathcal{K}}$ découle de sa symétrie. L'équation problématique Eq. (E.4) se lit alors comme suit :

$$-\boldsymbol{\nabla} \cdot \begin{bmatrix} K_{b_\parallel} & 0 & 0 \\ 0 & K_{b_\perp} & 0 \\ 0 & 0 & K_{b_\perp} \end{bmatrix} \begin{Bmatrix} \partial T/\partial b_\parallel \\ \partial T/\partial b_\perp^1 \\ \partial T/\partial b_\perp^2 \end{Bmatrix} + \mu T = S. \tag{E.5}$$

Dans ce qui suit, les méthodes basées sur des pochoirs indépendants du tenseur de diffusion et utilisant la formule de différenciation dans les directions x, y, et z seront dénotées *méthodes non alignées*. En revanche, les méthodes utilisant des pochoirs adaptés à la direction de $\boldsymbol{b}$ seront dénotées *méthodes alignées*.

E.3.3 Discrétisation numérique à l'intérieur du domaine

Le domaine de calcul est le cube $[0, 2\pi] \times [0, 2\pi] \times [0, 2\pi]$ dans les directions (x, y, z), respectivement. Il est considéré comme ouvert, la discrétisation des conditions limites à la frontière du domaine considéré par la suite.

E.3.4 Définition et notation de la grille

La grille est structurée et uniforme. Chaque cellule de la grille peut être traitée par des indices (i, j, k), et chaque sommet a des coordonnées $x_i = i(2\pi/N_x)$, $y_j = j(2\pi/N_y)$, $z_k = k(2\pi/N_z)$ for $(i, j, k) \in [1, N_x] \times [1, N_y] \times [1, N_z]$, où N_x, N_y, N_z sont les nombres de points dans chaque direction. Les distances entre les points de la grille sont définies comme $\Delta x = x_{i+1} - x_i$, $\Delta y = y_{j+1} - y_j$ et $\Delta z = z_{k+1} - z_k$. Pour clarté, (i, j, k) est également identifié par $\lambda = (i-1)\, N_y\, N_z + (j-1)\, N_z + k$ ($\lambda = (i-1)\, N_y + j$ en cas de $x - y$).

Dans ce qui suit, la discrétisation sera orientée, avec $\boldsymbol{b}$ définissant la direction positive locale à tout point (i, j, k). Les quantités à discréditer peuvent donc être remplacées par $+$ ou $-$. Dans ce qui suit, l'ensemble des valeurs aux points de la grille, et aux points où les flux sont estimés, sera désigné par l'espace de la grille (GS) et l'espace du flux (FS), respectivement. Toutes les quantités appartenant à FS seront identifiés par tilde $\widetilde{}$.

E.3.5 Discrétisation du gradient parallèle $\nabla_\parallel$

La discrétisation est faite conservative en utilisant une formulation de volume fini. En supposant ici $\nabla \cdot \boldsymbol{b} = 0$, la définition intégrale suivante de $\nabla_\parallel$ peut être utilisée pour chaque volume de contrôle K, de volume V et de surface S qui nous permet d'estimer le gradient parallèle du flux à S :

$$\nabla_{\|} T = \frac{\mathbf{b}}{\|\mathbf{b}\|} \cdot \nabla T = \frac{1}{\|\mathbf{b}\|} \nabla \cdot (T\boldsymbol{b}) = \lim_{V(K)\to 0} \frac{1}{\|\mathbf{b}\| V(K)} \int_{S(K)} T\boldsymbol{b} \cdot \boldsymbol{n} dS \tag{E.6}$$

Le volume de contrôle K autour de chaque point de grille (i, j, k) est défini par le polygone avec des coins $(i, j \pm \frac{1}{2}, k \pm \frac{1}{2})$ dans le $y - z$-plane $\mathcal{X}i$, et extrudé le long de la direction parallèle jusqu'aux plan $\mathcal{X}_{i\pm\frac{1}{2}}$ (Fig. E.4). A ces plans, il chevauche partiellement les volumes de contrôle voisins définis à partir des points de grille situés dans les plans adjacents $\mathcal{X}_{i+1}$ et $\mathcal{X}_{i-1}$. Dans ce qui suit, nous ne considérerons par simplicité que les volumes de contrôle voisins définis en $\mathcal{X}_{i+1}$, la discrétisation étant similaire pour les volumes de contrôle définis en $\mathcal{X}_{i-1}$.

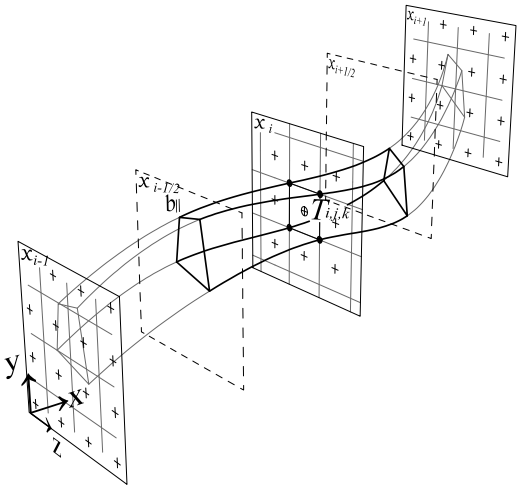


Figure E.4: Graphique d'un volume de contrôle (lignes en gras) défini dans GS autour de $T_{ijk} \in \mathcal{X}_i$, et entre les plans $\in \mathcal{X}_{i\pm\frac{1}{2}}$. Cas général avec $\boldsymbol{b}(x, y, z)$.

Les surfaces de contact entre les volumes de contrôle et ses voisins sont les suivantes: a_p, $p = 1, ..., N$, N étant le nombre total de zones de contact entre deux plans $\mathcal{X}$ adjacents (Fig. E.5a). Pour chaque surface de contact a_p, on considère la ligne qui passe par son barycenter bc_p et suit la direction parallèle, comme illustré dans la Fig. E.5b. Il intercepte les deux plans $\mathcal{X}_i$ et $\mathcal{X}_{i+1}$ à deux points de coordonnées (x_i, y^-, z^-) et (x_{i+1}, y^+, z^+), où $(y^\pm, z^\pm)$ sont définis entre $(x_{i+1/2})$ et (x_{i+1}) pour + (vers le sens positif des coordonnées, voir un croquis sur Fig. E.6) ou $(x_{i+1/2})$ et (x_i) pour − (vers le sens négatif des coordonnées) :

$$y^+ = y + \int_{x_{i+1/2}}^{x_{i+1}} \frac{b_y}{b_x} dx, \quad y^- = y + \int_{x_{i+1/2}}^{x_i} \frac{b_y}{b_x} dx, \tag{E.7}$$

où b_x, b_y les composantes de $\boldsymbol{b}$ dans le cadre cartésien (x, y, z). L'expression z^+ et z^- directions sont obtenues en remplaçant b_y par b_z dans Eq. E.7.

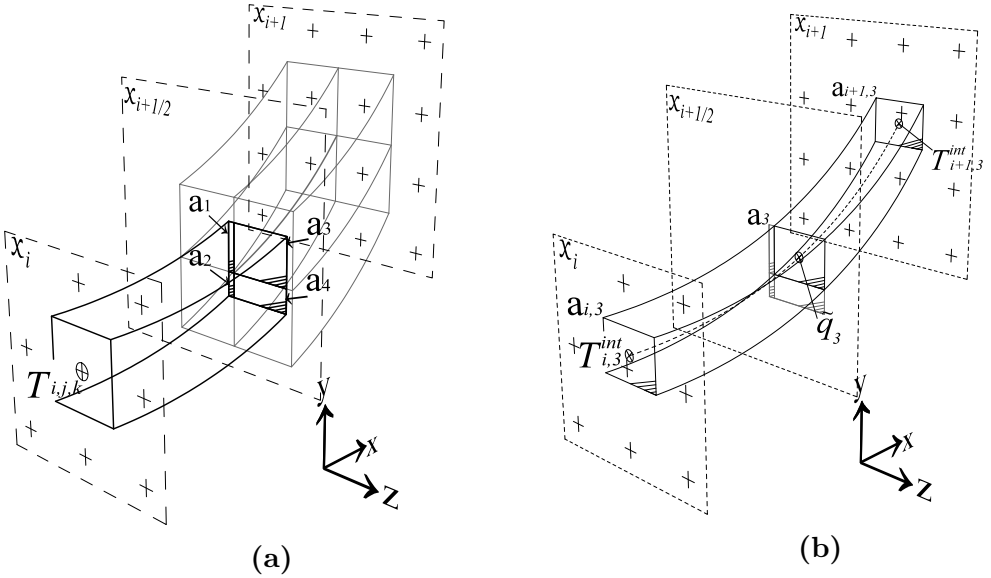


Figure E.5: Exemple de chevauchement des volumes de contrôle montrant les surfaces de contact a_p pour $p = 1, ..., 4$. (b) Chaque surface superposée permet de définir un volume de contrôle en FS, dénoté $\widetilde{\text{CV}}_p$. Les quantités utilisées pour évaluer le flux parallèle $\widetilde{q}_3$ à $\mathcal{X}_{i+1/2}$ à travers la surface spécifique a_3 sont incluses dans la figure. Ici, $\boldsymbol{b}(x)$ pour la simplicité.

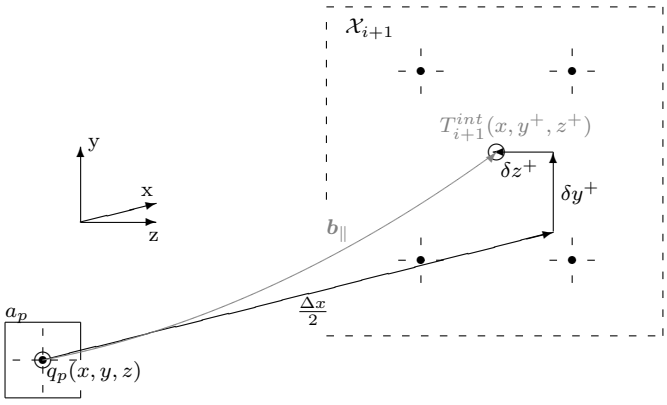


Figure E.6: Diagramme de suivi des vecteurs de diffusion parallèle de $(x, y, z) \in \mathcal{X}_{i+1/2}$ à $(x, y^+, z^+) \in \mathcal{X}_{i+1}$. δy^+ et δz^+ correspondent à la partie intégrale de l'Eq. E.7 pour l'indice $+$ et sa partie correspondante dans l'équation pour z^-, respectivement.

Les valeurs de champ à ces points sont obtenues par interpolation des valeurs de champ aux points environnants avec des fonctions $f^{int}_{i\,p}$ et $f^{int}_{i+1\,p}$ dans les plans correspondants. Nous pouvons écrire pour chaque p:

$$T^{int}_{i\,p} = f^{int}_{i\,p}\left(\{T_{i\,j\,k}\}_{i\,j=1,\dots N_y\,k=1,\dots N_z}\right), \tag{E.8}$$
$$T^{int}_{i+1\,p} = f^{int}_{i+1\,p}\left(\{T_{i+1\,j\,k}\}_{i\,j=1,\dots N_y\,k=1,\dots N_z}\right). \tag{E.9}$$

Ainsi, le gradient parallèle $\widetilde{(\nabla_{b\|}T)}_p$ peut être discrétisé en approchant l'Eq. E.6 comme suit :

$$\widetilde{(\nabla_{b\|}T)}_p = \frac{1}{\|\mathbf{b}\|\widetilde{\Delta V_p}}\left(T^{int}_{i+1\,p}\,a_{i+1\,p}\,\boldsymbol{b}_{i+1\,p}\cdot\boldsymbol{n}_{i+1\,p} + T^{int}_{i\,p}\,a_{i\,p}\,\boldsymbol{b}_{i\,p}\cdot\boldsymbol{n}_{i\,p}\right), \tag{E.10}$$

où $\widetilde{\Delta V_p}$ est le volume obtenu en intégrant la surface a_p le long de la direction parallèle entre $\mathcal{X}_i$ et $\mathcal{X}_{i+1}$, et $\boldsymbol{n}_p$ le vecteur normal vers la surface pertinente. La position à laquelle le flux est évalué est définie par le triplet $(\tilde{x}_p, \tilde{y}_p, \tilde{z}_p)$, qui sont les coordonnées du barycentre de la surface a_p.

La discrétisation du gradient parallèle (Eq. E.10) définit la transformation linéaire Q de l'espace des valeurs de grille (GS) dans l'espace des valeurs de flux (FS) :

$$\begin{aligned} Q: \quad \mathrm{GS} &\to \mathrm{FS} \\ \{T\} &\to \{\widetilde{\nabla_{b\|}T}\} \end{aligned} \tag{E.11}$$

afin que les valeurs de gradient soient données par :

$$\widetilde{(\nabla_{b\|}T)}_p = \sum_\lambda Q_{p\lambda}T_\lambda, \tag{E.12}$$

Il est important de noter ici que le calcul des contributions de gauche et de droite de l'Eq. E.10 sont construits pour satisfaire au niveau discret le fait que le flux de $\boldsymbol{b}$ sur toute la surface d'un volume fermé est nul, c.-à-d.:

$$a_{i\,p}\boldsymbol{b}_{i\,p}\cdot\boldsymbol{n}_{i\,p} + a_{i+1\,p}\boldsymbol{b}_{i+1\,p}\cdot\boldsymbol{n}_{i+1\,p} = 0 \tag{E.13}$$

Ceci garantit que Q est localement non puissant pour tout champ de température constante, c.-à-d. $\sum_\lambda Q_{p\lambda}T_\lambda = 0$ pour tous les p si T_λ est constant. Cette propriété est cruciale pour la conservation du régime. Il est également intéressant de noter que cela peut être réalisé en général par une discrétisation cohérente seulement si le champ vecteur $\boldsymbol{b}$ est de divergence zéro.

E.3.6 Discrétisation du Laplacien parallèle $\nabla\cdot(\mathbf{b}K_{b\|}\nabla_{b\|})$

L'expression pour le gradient parallèle ci-dessus permet maintenant d'utiliser la Méthode de l'Opérateur de Soutien (SOM) [MS08, SS94, MRS98, LMS14, MSS00] pour obtenir le

laplacien parallèle, comme on le trouve dans Stegmeir et al. [SCM+16] dans une diffusion hautement anisotrope. Pour toute fonction $T \in \mathcal{H}^2$, et $\Psi \in \mathcal{H}^1$, les deux continu en Ω, la formule de Green define:

$$\int_\Omega (\boldsymbol{\nabla} \cdot \mathbf{u}) \Psi dV + \int_\Omega \mathbf{u} \cdot \nabla \Psi dV = \int_\Gamma (\Psi \mathbf{u}) \cdot \mathbf{n} dS. \tag{E.14}$$

Considérant $\mathbf{u} = -\mathcal{K}, \nabla T$, la formule de Green relie le gradient et les opérateurs de divergence. Selon la définition du produit $\mathcal{L}^2$-inner dans le champ scalaire et les espaces vectoriels H et $\mathbf{H}$, respectivement, nous écrivons Eqs. E.15, E.16 pour chacun :

$$\langle -\nabla \cdot \mathcal{K} \nabla T, \Psi \rangle_H = -\int_\Omega \boldsymbol{\nabla} \cdot (\mathbf{u}\Psi) dV + \int_\Gamma (\Psi \mathbf{u}) \cdot \mathbf{n} dS. \tag{E.15}$$

$$\langle -\mathcal{K} \nabla T, \nabla \Psi \rangle_{\mathbf{H}} = \int_\Omega (\mathbf{u} \cdot \nabla \Psi) dV, \tag{E.16}$$

Le document de référence E.14 établit le lien entre les opérateurs de gradient et de divergence en tant que l'un est l'auto-adjoint de l'autre. En considérant des conditions limites appropriées (Dirichlet bi-périodique ou homogène), la formule de Green permet de définir l'opérateur de diffusion parallèle directement à partir du gradient parallèle comme :

$$\langle -\nabla \cdot (\mathbf{b} \mathcal{K} \nabla_{b\|} T), \Psi \rangle = \langle \mathcal{K} \nabla_{b\|} T, \nabla_{b\|} \Psi \rangle, \tag{E.17}$$

Même si Eq. E.17 est sans ambiguïté au niveau continu, il implique deux produits internes, l'un défini en GS, et l'autre en FS pour toutes les fonctions f et g comme :

$$\langle f, g \rangle_{\mathrm{GS}} = \sum_\lambda f_\lambda \, g_\lambda \, \Delta V_\lambda, \tag{E.18}$$

$$\langle f, g \rangle_{\mathrm{FS}} = \sum_p \widetilde{f}_p \, \widetilde{g}_p \, \widetilde{\Delta V_p}. \tag{E.19}$$

Selon le Eq. E.12, le produit intérne en FS peut être estimé au niveau discret en utilisant les évaluations de la diffusion sur les points de flux dénotés par $K_{b_\|, p}$ comme :

$$\langle [K] \nabla_{b\|} T, \nabla_{b\|} \Psi \rangle_{\mathrm{FS}} \approx \sum_p \widetilde{\Delta V_p} \Big(K_{b_\| \, p} \sum_\lambda Q_{p\,\lambda} T_\lambda \Big) \Big(\sum_\mu Q_{p\,\mu} \Psi_\mu \Big). \tag{E.20}$$

En fonction du nombre de surfaces de contact a_p, un certain nombre de valeurs de flux peut être associé pour chaque λ. En ce qui concerne le formalisme SOM [SS94], $Q_{p\lambda}$ de Eq. E.10, défini en FS est ici le *opérateur principal*. La discrétisation de la divergence (*opérateur derivé* en termes de SOM) est définie de FS en GS comme l'anti-adjoint de Q, obtenue par l'analogue discret de l'Eq. E.17. Ensuite, l'opérateur complet $(\nabla \cdot [K] \nabla_{b\|})$ est endomorphique en GS. Le côté gauche de l'Eq. E.17 mène au niveau discret à :

$$\langle \nabla \cdot [K] \nabla_{b\|} T, \Psi \rangle \approx \sum_{\sigma} (\nabla \cdot [K] \nabla_{b\|} T)_{\sigma} \Psi_{\sigma} \Delta V_{\sigma}. \tag{E.21}$$

Selon les Eqs. E.17, E.20 et E.21, on déduit par identification que :

$$-(\nabla \cdot [K] \nabla_{b\|} T)_{\lambda} \approx \frac{1}{\Delta V_{\lambda}} \sum_{p} \left(K_{b_{\|}\, p} Q_{p\lambda} \sum_{\mu} Q_{p\,\mu} T_{\mu} \widetilde{\Delta V_p} \right), \tag{E.22}$$

La somme sur le terme μ indique la construction des flux en FS de GS. Eq. E.22 mène :

$$(\nabla \cdot [K] \nabla_{b\|} T) \approx -\Delta V^{-1} Q^T \widetilde{[K]} \widetilde{\Delta V_p}\, Q\, T \tag{E.23}$$

Lors de la multiplication par le volume de la cellule ΔV, le SOM fournit une matrice discrète symétrique: le produit d'un opérateur et son anti-adjoint est un opérateur auto-adjoint négatif. La construction utilisée par SOM reflète cette symétrie et donne une matrice symétrique.

$$\mathcal{A}_{\lambda\mu} \Delta V_{\lambda} = \sum_{p} Q_{p\lambda} Q_{p\mu} \widetilde{[K]_p \Delta V_p} = \sum_{p} Q_{p\mu} Q_{p\lambda} \widetilde{[K]_p \Delta V_p} = \mathcal{A}_{\mu\lambda} \Delta V_{\mu}, \tag{E.24}$$

où $\widetilde{[K]_p} \widetilde{\Delta V_p} = \widetilde{[K \Delta V]}$ est une matrice carrée diagonale.

Finalement, la prudence du régime est vérifiée en prenant le cas particulier $\Psi = 1$:

$$-\langle \nabla \cdot [K] \nabla_{b\|} T, 1 \rangle_{\mathrm{GS}} = \langle [K] Q T, Q 1 \rangle_{\mathrm{FS}} = 0$$

ce qui veut dire que la moyenne du Laplacian parallèle sur le domaine de calcul est nulle. Cette propriété découle de Eq. E.13, ce qui implique $Q \cdot 1 = 0$, et implique conservativité :

$$\langle (1 - \nabla \cdot [K] \nabla_{b\|}) T, 1 \rangle_{\mathrm{GS}} = \langle T, 1 \rangle_{\mathrm{GS}}$$

Le schéma proposé préserve donc trois propriétés de l'opérateur continu, à savoir l'auto-adjoint, la positivité et la conservativité [SS94, LMS14].

E.3.7 Test numériques

Des tests numériques ont été effectués sur l'Eq. refHelmholtz˙1 en 2D dans le plan (x, y). Ainsi, la matrice de rotation $\boldsymbol{\mathcal{R}}$ de l'Eq. E.4 définit comme :

$$\boldsymbol{\mathcal{R}} = \begin{bmatrix} \cos\alpha & -\sin\alpha \\ \sin\alpha & \cos\alpha \end{bmatrix},$$

où l'angle α mesure le non-alignement entre le vecteur unitaire le long de la direction de l'anisotropie $\boldsymbol{b}$ et l'axe x. Ainsi, il résulte $\boldsymbol{b} = (\cos\alpha, \sin\alpha, 0)t$ (Fig. E.7). $\boldsymbol{b}$ sera supposé constant dans les tests.

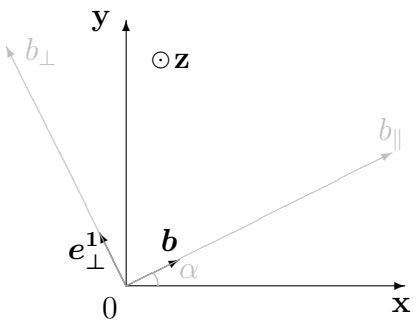


Figure E.7: Directions des principaux axes du tenseur de diffusion dans le plan (x, y). α définit l'angle de désalignement des axes principaux par rapport aux directions des points de grille.

Dans tous les tests $\mu = 1$ (voir une discussion sur l'influence de μ dans l'Appendix A).

Les méthodes de discrétisation suivantes ont été envisagées pour les essais :

- La méthode *classique* se référant à l'approche asymétrique [vEKdB14].
- La méthode *Günter* se référant à l'approche symétrique proposée par Günter *et al.* [GYKL05],
- La méthode *Ottaviani* se référant à une approche alignée (stencil orienté) basée sur une interpolation parallèle de second ordre polynomiale [Ott11, HO13].
- La méthode *Stegmeir* se référant à une approche alignée basée sur une interpolation linéaire [SCM+16].
- La méthode *présenté* se référant au travail accompli dans ce document. Il étend la méthode textitStegmeir à une discrétisation conservatrice dans les deux directions parallèles et perpendiculaires et à une discrétisation efficace de la condition limite dans les domaines délimités.

Les deux premières méthodes utilisaient des pochoirs indépendants du tenseur de diffusion, et se trouvent donc dans la classe des méthodes *non-alignées*. En revanche, les autres méthodes appartiennent à la classe *méthodes alignées*, telle que définie dans la section 2.

Détails numériques

Le terme source manufacturé S_a suivant a été considéré, correspondant à la superposition d'une constante, d'une contribution alignée et d'une contribution non alignée par rapport

$\mathbf{b}_\parallel$:

$$S_a(x,y) = C_1 + C_2 \cos\left(m_y y + m_{x,1} x\right) + C_3 \sin\left(m_{x,2} x\right) \tag{E.25}$$

Ce terme source correspond en effet à la superposition de fluctuations qui varient rapidement dans le sens perpendiculaire en étant uniformes le long du sens parallèle. L'angle $\phi = \tan -1(m_{x,1}/m_y)$ définit l'orientation des modes alignés. Les modes non alignés varient seulement en x. Dans le cas où $\phi = \alpha$, α étant l'angle d'orientation parallel, la résolution d'Eq. E.3.1 avec $K_{b\perp} = 0$ mène à la solution suivante :

$$T_a(x,y) = C_1 + C_2 \cos\left(m_y y + m_{x,1} x\right) + \frac{1}{1 + K_{b\parallel} m_{x,2}} C_3 \sin\left(m_{x,2} x\right) \tag{E.26}$$

Les fluctuations liées au premier terme devraient alors dominantes, et le coefficient d'amortissement de cette contribution particulière est un bon indicateur de la qualité de la discrétisation utilisée.

Tous les tests de précision ont été effectués avec $\alpha = \tan^{-1}(4/27)$, $m_y = 27$, $m_{x,1} = 4$, $m_{x,2} = 2$, $C_1 = 0$, $C_2 = 1$, $C_3 = 0.25$. Les champs 2D de T_a sont affichés sur Fig. E.8 pour cas peu ($K_{b_\parallel} = K_{b_\perp}$) et très ($K_{b_\parallel} = 10^6 K_{b_\perp}$) anisotropes, montrant ou non des modulations parallèles dans la direction parallèle, respectivement.

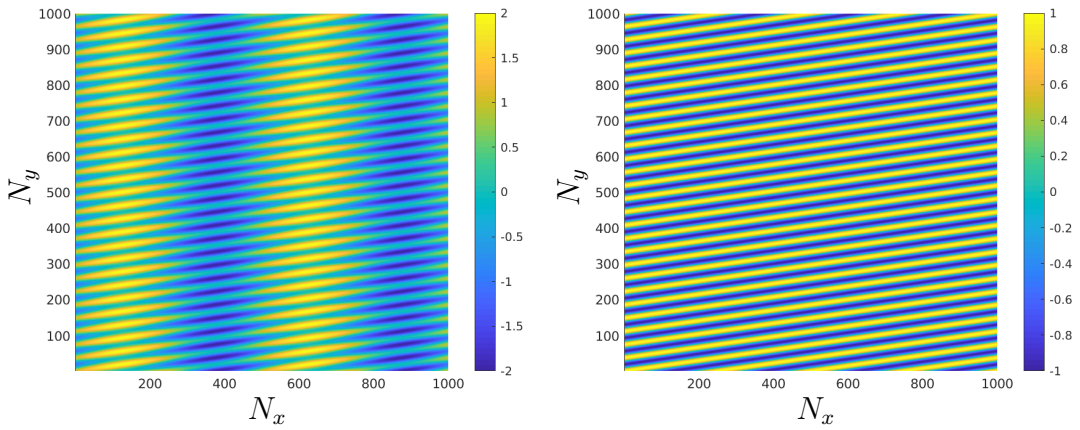


Figure E.8: Graphiques 2D de T_a lors de la résolution d'un Eq.˙ refHelmholtz analytique avec le terme source S_a Eq. E.25 for (a) an isotropic case $K_{b_\parallel} = K_{b_\perp}$ and (b) a strongly anisotropic case $K_{b_\parallel} = 10^6 K_{b_\perp}$. Le terme source S_a a les valeurs de paramètres suivantes: $C_1 = 0$, $C_2 = 1$, $C_3 = 1$, $m_y = 27$, $m_{x,1} = 4$ and $m_{x,2} = 2$.

Pour un point de vue pratique et les utilisateurs de codes, il est très utile de déterminer quelle est la résolution nécessaire en fonction de la précision ciblée, la diffusion parallèle

et la méthode de discrétisation choisie. Nous montrons des graphiques 2D des erreurs donnés par la norme $\mathcal{H}1$ en fonction de N_x et N_y et pour $K_{b\|} = 1$ et $K_{b\|} = 10^6$ qui correspondent à des solutions pour lesquelles les fluctuations non alignées sont faiblement ou fortement amorties, respectivement.

Pour les deux méthodes *non alignées*, les graphiques 2D sont affichés sur Figs. E.9, E.10, pour les méthodes *Classique* et *Günter* respectivement. Pour une faible diffusion parallèle, $K_{b\|} = 1$, les résultats montrent que la convergence nécessite une résolution minimale dans la direction x-(N_x environ $16-32$). Cependant, pour une grande diffusion parallèle, $K_{b\|} = 10^6$, la méthode *Classique* ne converge vers la solution pour aucune résolution testée. Pour la méthode *Günter*, le domaine de convergence est très petit, et la méthode ne converge qu'à haute résolution, à partir de 128×1024. Ensuite, les méthodes non alignées nécessitent de très grandes résolutions dans les deux directions $x - y$ lors de la résolution de cas hautement anisotropes.

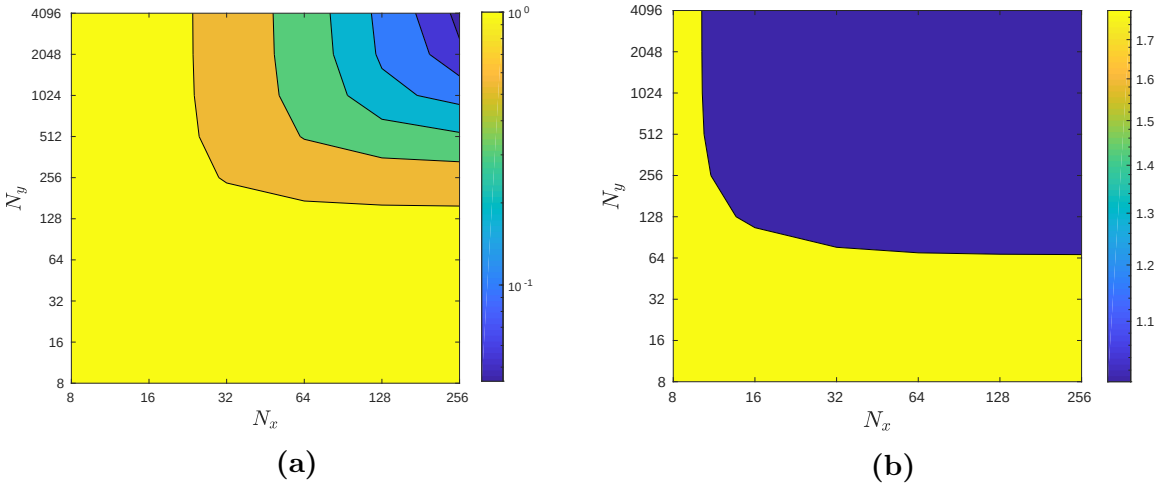
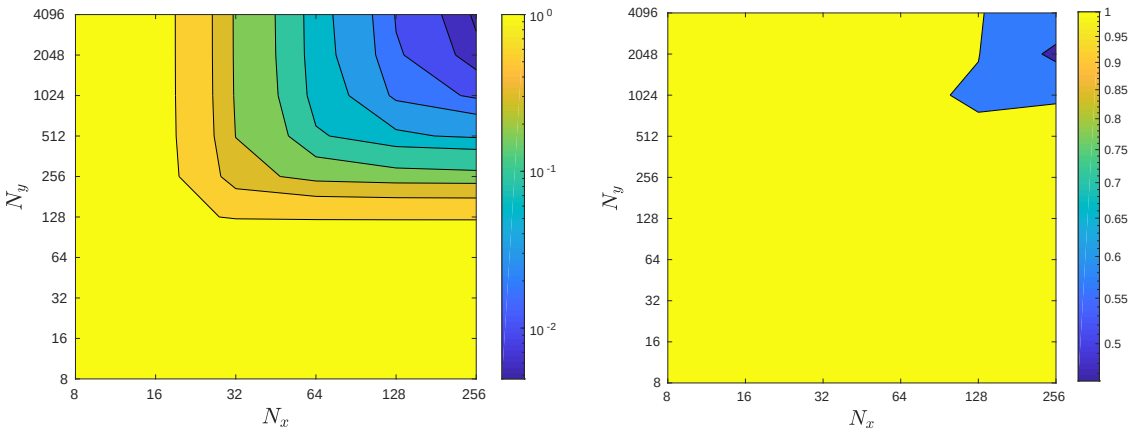


Figure E.9: Graphiques 2D de l'erreur $\mathcal{H}^1$-error en fonction de N_x, et N_y pour le *méthode Classique* avec $K_{b\perp} = 0$, (a) $K_{b\|} = 1$ et (b) $K_{b\|} = 106$. Le domaine informatique est bi-périodique.

Pour les trois méthodes de *alignées*, c.-à-d. les méthodes *Ottaviani, Stegmeir* et *présenté*, les résultats sont présentés dans les Figs. E.11, E.12, E.13, respectivement. Dès que la résolution dans le y-direction est suffisante, les méthodes convergent avec seulement quelques points dans le x-direction. À faible diffusion parallèle $K_{b\|} = 1$, la précision n'est que faiblement sensible à N_x et une grande amélioration de la précision peut être obtenue en augmentant seulement N_y.

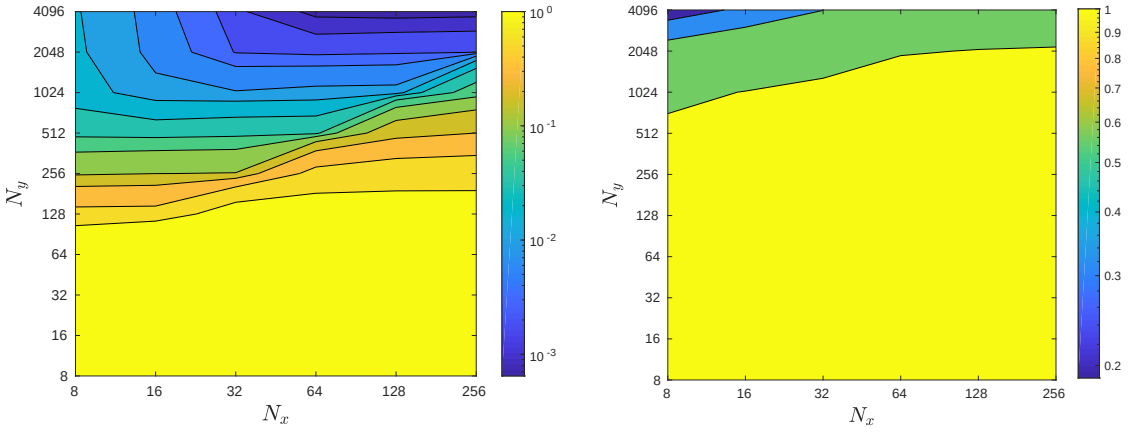
A large diffusion parallèle $K_{b\|} = 10^6$, les résultats montrent la supériorité des méthodes alignées par rapport aux méthodes non alignées. La diffusion numérique est réduite ici avec N_x, montrant une convergence rapide de $N_y \geq 512$. A haute résolution dans la direction x, les méthodes alignées ne convergent plus en raison de l'erreur accumulé



(a) **(b)**

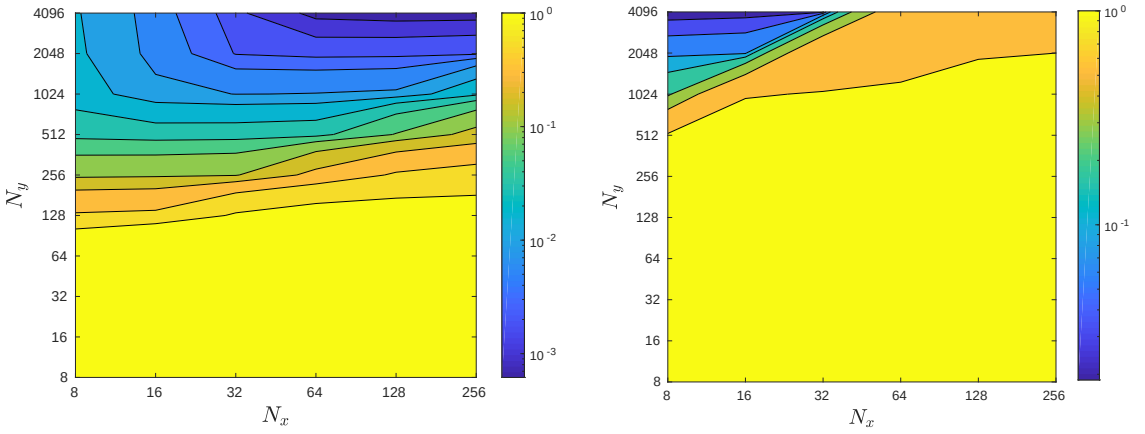
Figure E.10: Graphiques 2D de l'erreur $\mathcal{H}^1$-error en fonction de N_x, et N_y pour le *méthode Günter* avec $K_{b\perp} = 0$, (a) $K_{b\parallel} = 1$ et (b) $K_{b\parallel} = 10^6$. Le domaine informatique est bi-périodique.

d'interpolation dans la direction y. Ce dernier a un grand impact sur les différences finies quand Δx (qui est dans le dénominateur de différence finie) devient petit, étant cet effet amplifié autrement par la valeur de $K_{b\parallel}$.



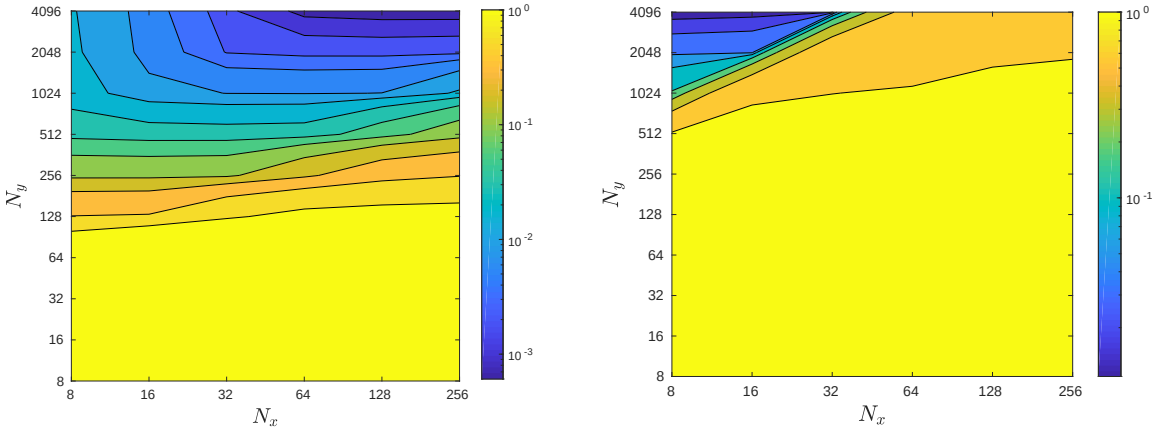
(a) **(b)**

Figure E.11: Graphiques 2D de l'erreur $\mathcal{H}^1$-error en fonction de N_x, et N_y pour le *méthode Ottaviani* avec $K_{b\perp} = 0$, (a) $K_{b\parallel} = 1$ et (b) $K_{b\parallel} = 106$. Le domaine informatique est bi-périodique.



(a) **(b)**

Figure E.12: Graphiques 2D de l'erreur $\mathcal{H}^1$-error en fonction de N_x, et N_y pour le *méthode Stegmeir* avec $K_{b\perp} = 0$, (a) $K_{b\parallel} = 1$ et (b) $K_{b\parallel} = 106$. Le domaine informatique est bi-périodique.



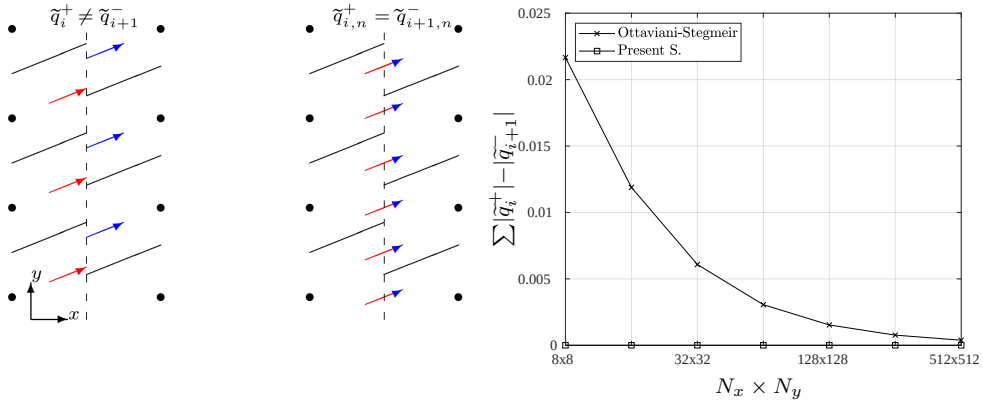
(a) **(b)**

Figure E.13: Graphiques 2D de l'erreur $\mathcal{H}^1$-error en fonction de N_x, et N_y pour le *méthode actuelle* avec $K_{b\perp} = 0$, (a) $K_{b\parallel} = 1$ et (b) $K_{b\parallel} = 106$. Le domaine informatique est bi-périodique.

Tests de conservativité

Une nouvelle caractéristique de la *méthode actuelle* par rapport aux *méthodes alignées* existantes de la littérature est d'impliquer une discrétisation conservative des flux pour l'opérateur parallèle.

Les *méthodes alignés* dans la littérature, (méthodes *Ottaviani* et *Stegmeir*) évaluent les flux au centre des faces de CV pour chaque plan $\mathcal{X}_i$. Cela conduit à un désalignement



(a) **(b)** **(c)**

Figure E.14: Graphiques montrant la discrétisation des flux entre les volumes de contrôle adjacents pour *méthode d'Ottaviani* (a) alignés sur les points de grille [Ott11] et (b) le *méthode actuelle* centré sur la surface de contact entre les volumes de contrôle. (c) Tracé pour différentes grilles $N_x \times N_y$ de la différence relative $\sum|\widetilde{q}_i^+|-|\widetilde{q}_{i+1}^-|$ entre les flux avant de $\mathcal{X}_i$ plan (flux rouges dans la graphique (a)) et les flux arrière de $\mathcal{X}_{i+1}$ avion (flux bleus dans la graphique (a)).

des flux entre les CV adjacents (Fig. E.14a) et donc à une méthode non conservative. La discrétisation des flux calculés au centre des faces communes de deux CV adjacents assure la conservativité de la *méthode actuelle*, Fig. E.14b. Cette définition du flux conduit à une définition symétrique des flux entre $mathcal X$ plans indépendamment de $K_{b\|}$. Pour montrer cela, un test a été effectué en considérant le terme source $T_a = 2 + sin(x)sin(y)$ avec un $K_{b\|} = 2 + sin(x)sin(y)$.

Le test considère différents N_{dof}, montrant la quantité $|\widetilde{q}_i^+|-|\widetilde{q}_{i+1}^-|$, représentant l'équilibre entre les flux rouges et bleus représentés dans Fig. E.14a pour textitOttaviani et textitMéthodes de Stegmeir (les deux schémas mènent la même définition de flux ici; voir [SCM+16]) et Fig. E.14b pour la méthode présenté. Les résultats du test, Fig. E.14c, montrent comment la définition symétrique et unique des flux de la méthode présenté conduit à un équilibre de flux parfait égal à zéro, ce qui signifie la différence de la quantité $|\widetilde{q}_i^+|-|\widetilde{q}_{i+1}^-|$ est toujours égal à zéro par construction. Pour *Ottaviani-Stegmeir* $|\widetilde{q}_i^+|-|\widetilde{q}_{i+1}^-|$ la définition non symétrique des flux conduit à une somme déséquilibrée, qui augmente pour N_{dof} réduits.

Appendix F

Proof of the accepted article in Journal of Computational Physics

Contents lists available at ScienceDirect

Journal of Computational Physics

www.elsevier.com/locate/jcp

A new conservative finite-difference scheme for anisotropic elliptic problems in bounded domain

J.A. Soler [a], F. Schwander [a], G. Giorgiani [c], J. Liandrat [b], P. Tamain [c], E. Serre [a]

[a] Aix Marseille Univ., CNRS, Centrale Marseille, M2P2, Marseille, France
[b] Aix Marseille Univ., CNRS, Centrale Marseille, I2M, Marseille, France
[c] IRFM, CEA Cadarache, F-13108 St. Paul-lez-Durance, France

ARTICLE INFO

Article history:
Received 21 January 2019
Received in revised form 30 October 2019
Accepted 31 October 2019
Available online xxxx

Keywords:
Anisotropic operators
Conservative finite-difference scheme
Aligned interpolation

ABSTRACT

Highly anisotropic elliptic problems occur in many physical models that need to be solved numerically. A direction of dominant diffusion is thus introduced (called here parallel direction) along which the diffusion coefficient is several orders larger of magnitude than in the perpendicular one. In this case, finite-difference methods based on misaligned stencils are generally not designed to provide an optimal discretization, and may lead the perpendicular diffusion to be polluted by the numerical error in approximating the parallel diffusion.

This paper proposes an original scheme using non-aligned Cartesian grids and interpolations aligned along a parallel diffusion direction. Here, this direction is assumed to be supported by a divergence-free vector field which never vanishes and it is supposed to be stationary in time. Based on the Support Operator Method (SOM), the self-adjointness property of the parallel diffusion operator is maintained on the discrete level. Compared with existing methods, the present formulation further guarantees the conservativity of the fluxes in both parallel and perpendicular directions. In addition, when the flow intercepts a boundary in the parallel direction, an accurate discretization of the boundary condition is presented that avoids the uncertainties of extrapolated far ghost points classically used and ensures a better accuracy of the solution. Numerical tests based on manufactured solutions show the method is able to provide accurate and stable numerical approximations in both periodic and bounded domains with a drastically reduced number of degrees of freedom with respect to non-aligned approaches.

1. Introduction

Elliptic partial differential systems are ubiquitous in physical models and numerical simulations. They occur in fluid models used in mechanics, geophysics, plasma physics, but also in other fields of research as in microelectronics, optics, image processing, and so on, the list being not exhaustive.

A typical problem to solve with appropriate boundary conditions is Poisson's equation that writes as:

$$-\nabla \cdot (\mathcal{K} \cdot \nabla) T = S, \quad \in \Omega \subset \mathbb{R}^3$$

E-mail address: eric.serre@univ-amu.fr (E. Serre).

https://doi.org/10.1016/j.jcp.2019.109093
0021-9991/

where $\mathcal{K}$ is the diffusion tensor. In many configurations the isotropy of the problem can be broken, and a preferred direction (direction of anisotropy) is thus introduced that leads to an anisotropic diffusion tensor. Since this work is motivated by the simulation of fusion plasmas in tokamak [1,2], the anisotropy direction is related to the anisotropy of magnetic field components in the tokamak and it is denoted as the parallel direction, with reference to the direction along the magnetic field lines. Assuming moreover that the diffusion is isotropic in the perpendicular directions, leads to the following expression of the tensor in a coordinate system whose axes coincide with the principal directions:

$$\begin{bmatrix} K_{b_\parallel} & 0 & 0 \\ 0 & K_{b_\perp} & 0 \\ 0 & 0 & K_{b_\perp} \end{bmatrix}$$

where $K_{b_\parallel}$ and $K_{b_\perp}$ are functions of space, with $K_{b_\parallel}/K_{b_\perp} \gg 1$.

Regular, structured mesh methods for conservation laws are generally not designed to discretize such anisotropic operators in an optimal way. They inherently favor directions aligned with mesh points, which can introduce systematic errors, even at high resolution, when the principal directions of the diffusion tensor are not aligned with the grid axes. The discretization can in particular produce a significant spurious numerical diffusion in the direction orthogonal to the anisotropy direction, which can significantly impact the perpendicular dynamics [3]. As recalled in van Es et al. [4], other problems can also arise as non-positivity near high gradients [5], stagnation or even loss of the convergence of the discretization error with mesh refinement, see [6,7]. Such spurious effects can be reduced by using a high-order scheme to discretize operators in the direction of anisotropy, but larger stencils are then required.

Many mathematical and numerical works have been devoted to the discretization of anisotropic diffusion operators (see a quite exhaustive list of references in van Es et al. [4]). More specifically in the frame of finite-difference methods, support operator methods (SOM) [8,9], also known as Mimetic finite-difference [10], allow to preserve at the discrete level the property that the negative divergence and gradient operators are mutually adjoint, so as to enforce the positive definiteness of the continuum problem when discretizing second-order partial differential equations. Hyman et al. [10] set the conditions to define Dirichlet, Neumann and Robin boundary conditions preserving SOM properties. In Morel et al. [11], a new version of SOM was proposed, which is said local because yielding a sparse diffusion matrix, in contrast to the traditional SOM which yields to a dense matrix representation.

In magnetized fusion, the intensity of the toroidal component of the magnetic field in the tokamak is much larger than in the poloidal one, leading to $K_{b_\parallel} \gg K_{b_\perp}$. The anisotropy direction is thus supported by a divergence-free vector field which never vanishes in this magnetic configuration. For such simulations, Günter et al. [7] proposed a finite-difference method with SOM conditions on rectangular grids, using a conservative approach in which fluxes are discretized on the dual mesh. Thanks to the discretization of the parallel operator, the spurious perpendicular diffusion observed in standard finite difference simulations for large anisotropies ($10^9 \leq K_{b_\parallel}/K_{b_\perp} \leq 10^{12}$) is strongly reduced. Other proposals have been made with the incentive of reducing the number of grid points when the system exhibits very strong diffusion in one specific direction. First magnetic-field-aligned approaches appeared in Refs. [12–15]. In these, the mesh is constructed so as to be aligned with the magnetic field. Scott [16] relieved the constraint of aligning the mesh by introducing a local coordinate system on each grid point. This approach avoids grid deformations by using the so-called "shifted metric" procedure, taking into account the Hamada [17] flux tube approach (global aligned coordinates), and splitting it into local shifted tubes related to grid points. Later on, Ottaviani [18] and Hariri & Ottaviani [19] introduced the Flux-Coordinate Independent approach, a field-aligned approach in non-aligned Cartesian and polar grids. Based on an interpolation along the parallel diffusion direction, this method is able to reduce the number of grid points in rectangular and cylindrical grids, reducing drastically the degrees of freedom of the computation for a given accuracy. The numerical diffusion in the perpendicular direction due to the discretization of the parallel operator is shown to decrease, and to become smaller than the one introduced by the classical Arakawa scheme used for advection terms [20]. In Stegmeir et al. [21,22] the Field Line Map (FLM) approach is presented. Based on the magnetic field lines trace geometry, Stegmeir and co-authors present an integration method for the gradient operator from interpolated field values in the magnetic field lines, before deriving the full diffusion operator with SOM. The resulting approach presents lower numerical diffusion than the physical perpendicular diffusion given by the Arakawa discretization, higher convergence tendency and a reduction of the number of unknowns as in [18]. Finally van Es et al. [4] compare the accuracy of Günter's scheme with field lines tracking approaches in regular Cartesian grids. The comparison is made using anisotropic test-cases and co-located grids, and it shows the good properties of the aligned approaches in terms of condition numbers of the resulting linear system and accuracy of the solution. Methods based on the elliptic problem reformulation into an equivalent one like Asymptotic Preserving methods [23,24] are not compared here. In this class of methods, Del Castillo-Negrete and Chacón [25] introduced a method which avoids the discrete matrix inversion (usually ill-conditioned for highly anisotropic diffusion tensors) towards the Lagrangian Green's function.

In this work, we propose a new finite-difference scheme based on interpolations aligned along the direction of the anisotropy corresponding to the direction defined by the dominant diffusion. The scheme is proved to be robust and conservative, and able to deal accurately with various boundary conditions. The paper is organized as follows: the mathematical model is presented in Sec. 2. The numerical scheme is presented in terms of conservative discretizations of the parallel and perpendicular operators in the interior of the domain in Sec. 3. The construction of the stencils in a 2D domain is illustrated for both curved and straight parallel field lines and possibly non uniform parallel diffusion in Sec. 4. This latest

Please cite this article in press as: J.A. Soler et al., A new conservative finite-difference scheme for anisotropic elliptic problems in bounded domain, J. Comput. Phys. (2019), https://doi.org/10.1016/j.jcp.2019.109093

also includes a new compatible approach to deal with bounded problems in Sec. 4.3. Finally, the accuracy, the conservativity and the efficiency in terms of grid resolution of the new scheme are then presented in Sec. 5 for 2D elliptic problems in both bi-periodic and bounded domain. The case of the Poisson's equation of prime importance in many applications is also included by considering general Robin boundary conditions. Results are analyzed with respect to existing aligned and non-aligned methods of the literature.

2. Mathematical model

Our focus lies in the resolution of strongly anisotropic diffusion problems using first-order, implicit time discretization, or anisotropic Poisson's equations occurring when investigating stationary solutions of the same diffusion problems. These two problems can be described by the following, generic, elliptic boundary value problem:

$$\begin{cases} -\nabla \cdot \mathcal{K} \nabla T + \mu T = S & \text{in } \Omega, \\ \beta \nabla_{b\parallel} T + \gamma T = g & \text{on } \Gamma, \end{cases} \tag{1}$$

where Ω is a bounded domain in $\mathbb{R}^3$ with boundary Γ, provided with an orthonormal basis $(\mathbf{e}_x, \mathbf{e}_y, \mathbf{e}_z)$ associated to Cartesian coordinates (x, y, z). We assume that the variables of the problem satisfy the usual ellipticity and regularity assumptions. μ is a positive (or zero) constant, and S is a given source term. The coefficients β and γ, and g are used to define general boundary conditions, that can be of Dirichlet ($\beta = 0$), Neumann ($\gamma = 0$) or Robin ($\beta \neq 0, \gamma \neq 0$) type. With such a problem, both periodic and bounded domains can be considered that allows us to disconnect the discretization of the equation in the interior of the domain and at the boundaries.

The anisotropy of the problem is taken into account via the definition of the symmetric diffusion tensor $\mathcal{K}$. The first eigenvalue $K_{b_\parallel}$ is assumed to fix the dominant diffusion direction ($K_{b_\parallel} \gg K_{b_\perp}$), that we can identify with the normalized eigenvector $\mathbf{b}$. $\boldsymbol{b}$ is assumed here to be divergence-free and never vanishing. Under these assumptions, the system reads:

$$-\nabla \cdot \left[\mathcal{R} \begin{bmatrix} K_{b_\parallel} & 0 & 0 \\ 0 & K_{b_\perp} & 0 \\ 0 & 0 & K_{b_\perp} \end{bmatrix} \mathcal{R}^{-1} \right] \begin{Bmatrix} \partial T/\partial x \\ \partial T/\partial y \\ \partial T/\partial z \end{Bmatrix} + \mu T = S. \tag{2}$$

where $\mathcal{R}$ defines a 3D rotation matrix.

Gradients along the parallel and perpendicular directions are then defined as $\nabla_{b\parallel} = \boldsymbol{b} \cdot \nabla$ and $\nabla_{b\perp} = \nabla - \boldsymbol{b} \nabla_{b\parallel}$, respectively.

Great simplifications can be obtained by defining an orthonormal basis constituted by the normalized eigenvectors of $\mathcal{K}$, namely $(\mathbf{b}, \mathbf{e}_\perp^1, \mathbf{e}_\perp^2)$, and the associated aligned coordinate system $(b_\parallel, b_\perp^1, b_\perp^2)$. The orthogonality of the eigenvectors of $\mathcal{K}$ follows from its symmetry. The problem equation Eq.(2) then reads:

$$-\nabla \cdot \begin{bmatrix} K_{b_\parallel} & 0 & 0 \\ 0 & K_{b_\perp} & 0 \\ 0 & 0 & K_{b_\perp} \end{bmatrix} \begin{Bmatrix} \partial T/\partial b_\parallel \\ \partial T/\partial b_\perp^1 \\ \partial T/\partial b_\perp^2 \end{Bmatrix} + \mu T = S. \tag{3}$$

In the following, the methods based on stencils independent of the diffusion tensor and using differentiation formula in the x, y, and z directions will be denoted *non-aligned methods*. In contrast, the methods using stencils adapted to the direction of $\boldsymbol{b}$ will be denoted *aligned methods*.

3. Numerical discretization in the interior of the domain

The computational domain is the cube $[0, 2\pi] \times [0, 2\pi] \times [0, 2\pi]$ in the (x, y, z) directions, respectively. It is considered as open, the discretization of the boundary conditions at the border of the domain being considered thereafter.

3.1. Grid definition and notation

The grid is structured and uniform. Each cell in the grid can be addressed by indices (i, j, k), and each vertex has coordinates $x_i = i(2\pi/N_x)$, $y_j = j(2\pi/N_y)$, $z_k = k(2\pi/N_z)$ for $(i, j, k) \in [1, N_x] \times [1, N_y] \times [1, N_z]$, where N_x, N_y, N_z are the numbers of points in each direction. Distances between grid points are defined as $\Delta x = x_{i+1} - x_i$, $\Delta y = y_{j+1} - y_j$ and $\Delta z = z_{k+1} - z_k$. For clarity, (i, j, k) is also labeled by $\lambda = (i-1) N_y N_z + (j-1) N_z + k$ ($\lambda = (i-1) N_y + j$ in $x - y$ cases).

In the following, the discretization will be oriented, with $\boldsymbol{b}$ defining the local positive direction at any (i, j, k) point. Quantities to discretize may thus eventually be superscripted with $+$ or $-$ when needed. In the following, the set of values at the grid points, and at the points where fluxes are estimated will be denoted grid space (GS) and flux space (FS), respectively. All quantities belonging to FS will be superscripted by tilde $\tilde{}$.

Please cite this article in press as: J.A. Soler et al., A new conservative finite-difference scheme for anisotropic elliptic problems in bounded domain, J. Comput. Phys. (2019), https://doi.org/10.1016/j.jcp.2019.109093

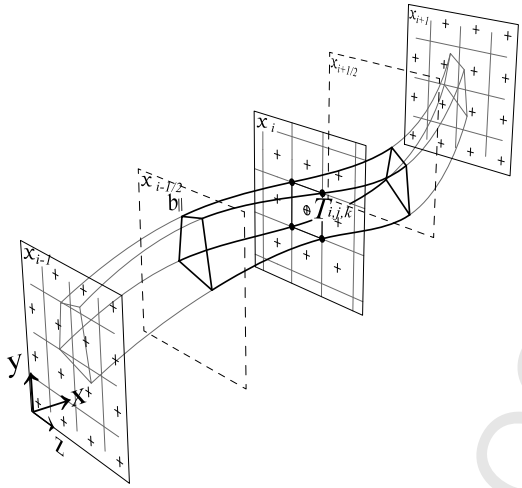


Fig. 1. Sketch of a control volume (bold lines) defined in the grid space GS around $T_{i\,j\,k} \in \mathcal{X}_i$, and between $\mathcal{X}_{i\pm\frac{1}{2}}$ planes. General case with $\boldsymbol{b}(x, y, z)$.

3.2. Discretization of the parallel gradient $\nabla_\parallel$

The discretization is made conservative by using a finite-volume formulation. Assuming here $\nabla \cdot \boldsymbol{b} = 0$, the following integral definition of $\nabla_\parallel$ can be used for each control volume K, of volume V and surface S that allows us to estimate the parallel gradient from the flux through S:

$$\nabla_\parallel T = \frac{\mathbf{b}}{\|\mathbf{b}\|} \cdot \nabla T = \frac{1}{\|\mathbf{b}\|} \nabla \cdot (T\boldsymbol{b}) = \lim_{V(K)\to 0} \frac{1}{\|\mathbf{b}\| V(K)} \int\limits_{S(K)} T\boldsymbol{b} \cdot \boldsymbol{n} dS \tag{4}$$

The control volume K around each grid point (i, j, k) is defined by the polygon with corners $(i, j \pm \frac{1}{2}, k \pm \frac{1}{2})$ in the y–z-plane $\mathcal{X}_i$, and extruded along the parallel direction up to the planes $\mathcal{X}_{i\pm\frac{1}{2}}$ (Fig. 1). At these planes, it partially overlaps neighboring control volumes defined from grid points located in the adjacent planes $\mathcal{X}_{i+1}$ and $\mathcal{X}_{i-1}$. In the following, we will only consider by simplicity neighboring control volumes defined in $\mathcal{X}_{i+1}$, the discretization being similar for control volumes defined in $\mathcal{X}_{i-1}$.

The contact surfaces between control volumes and its neighbors are denoted a_p, $p = 1, ..., N$, N being the total number of contact areas between two adjacent $\mathcal{X}$ planes (Fig. 2a). For each contact surface a_p, we consider the line that passes by its barycenter bc_p and follows the parallel direction, as illustrated in Fig. 2b. It intercepts the two planes $\mathcal{X}_i$ and $\mathcal{X}_{i+1}$ at two points of coordinates (x_i, y^-, z^-) and (x_{i+1}, y^+, z^+), where $(y^\pm, z^\pm)$ are defined between $(x_{i+1/2})$ and (x_{i+1}) for $+$ (in the positive sense of the coordinates, see a sketch on Fig. 3) or $(x_{i+1/2})$ and (x_i) for $-$ (in the negative sense of the coordinates):

$$y^+ = y + \int\limits_{x_{i+1/2}}^{x_{i+1}} \frac{b_y}{b_x} dx, \quad y^- = y + \int\limits_{x_{i+1/2}}^{x_i} \frac{b_y}{b_x} dx, \tag{5}$$

being b_x, b_y the components of $\boldsymbol{b}$ in the Cartesian frame (x, y, z). The expression for z^+ and z^- directions are obtained replacing b_y by b_z in Eq. (5).

The field values at these points are obtained by interpolation of the field values at the surrounding points with functions $f^{int}_{i\,p}$ and $f^{int}_{i+1\,p}$ in the corresponding planes. We can write for each p :

$$T^{int}_{i\,p} = f^{int}_{i\,p} \left(\{T_{i\,j\,k}\}_{i\,j=1,...N_y\,k=1,...N_z} \right), \tag{6}$$

$$T^{int}_{i+1\,p} = f^{int}_{i+1\,p} \left(\{T_{i+1\,j\,k}\}_{i\,j=1,...N_y\,k=1,...N_z} \right). \tag{7}$$

Thus, the parallel gradient $\widetilde{(\nabla_{b\parallel} T)_p}$ can be discretized by approximating Eq. (4) as follows:

$$\widetilde{(\nabla_{b\parallel} T)_p} = \frac{1}{\|\mathbf{b}\| \widetilde{\Delta V_p}} \left(T^{int}_{i+1\,p}\, a_{i+1\,p}\, \boldsymbol{b}_{i+1\,p} \cdot \boldsymbol{n}_{i+1\,p} + T^{int}_{i\,p}\, a_{i\,p}\, \boldsymbol{b}_{i\,p} \cdot \boldsymbol{n}_{i\,p} \right), \tag{8}$$

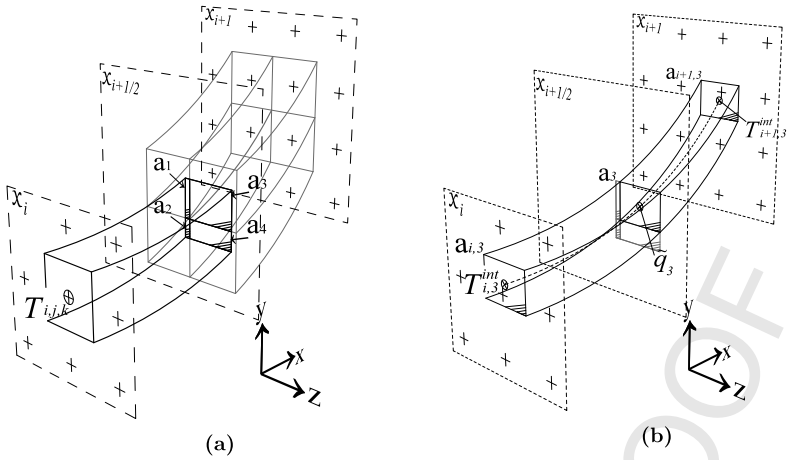


Fig. 2. (a) Example of control volumes overlapping showing the contact surfaces a_p for $p = 1, ..., 4$ (see Fig. 4 for a 2D version of the overlap). (b) Each overlapped surface allows to define a control volume in the flux space FS, denoted $\widetilde{\mathrm{CV}}_p$. Quantities used to evaluate the parallel flux $\widetilde{q}_3$ at $\mathcal{X}_{i+1/2}$ through the specific surface a_3 are included in the figure (see also Fig. 5). Here, $\boldsymbol{b}(x)$ for simplicity.

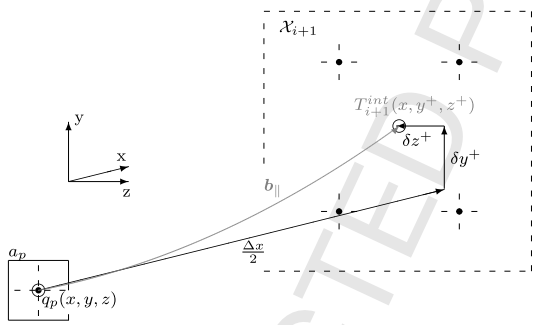


Fig. 3. Parallel diffusion vector tracking chart from $(x, y, z) \in \mathcal{X}_{i+1/2}$ to $(x, y^+, z^+) \in \mathcal{X}_{i+1}$. δy^+ and δz^+ correspond to the integral part of Eq. (5) for $+$ index, and its corresponding part in the equation for z^+, respectively.

where $\widetilde{\Delta V_p}$ is the volume obtained by integrating the surface a_p along the parallel direction between $\mathcal{X}_i$ and $\mathcal{X}_{i+1}$, and $\boldsymbol{n}_p$ the normal vector to the relevant surface. It is convenient to define a position at which the flux is evaluated, defined by the triplet $(\tilde{x}_p, \tilde{y}_p, \tilde{z}_p)$ which are the coordinates of the barycenter of the surface a_p.

The discretization of the parallel gradient (Eq. (8)) defines a linear map Q from the space of grid values (GS) into the space of flux values (FS):

$$\begin{aligned} Q &: \mathrm{GS} \to \mathrm{FS} \\ &\{T\} \to \{\widetilde{\nabla_{b\parallel} T}\} \end{aligned} \tag{9}$$

so that gradient values are given by:

$$(\widetilde{\nabla_{b\parallel} T})_p = \sum_\lambda Q_{p\lambda} T_\lambda, \tag{10}$$

It is important to note here that the computation of left and right contributions of Eq. (8) are constructed to satisfy at the discrete level the fact the flux of $\boldsymbol{b}$ across the surface of any closed volume is zero, i.e.:

$$a_{i\,p} \boldsymbol{b}_{i\,p} \cdot \boldsymbol{n}_{i\,p} + a_{i+1\,p} \boldsymbol{b}_{i+1\,p} \cdot \boldsymbol{n}_{i+1\,p} = 0 \tag{11}$$

This ensures that Q is locally nilpotent for any constant temperature field, i.e. $\sum_\lambda Q_{p\lambda} T_\lambda = 0$ for all p if T_λ is constant. This property is crucial to the conservativity of the scheme. It is also noteworthy that this can be achieved in general by a consistent discretization only if the vector field $\boldsymbol{b}$ is divergence-free.

Please cite this article in press as: J.A. Soler et al., A new conservative finite-difference scheme for anisotropic elliptic problems in bounded domain, J. Comput. Phys. (2019), https://doi.org/10.1016/j.jcp.2019.109093

3.3. Discretization of the parallel Laplacian $\nabla \cdot (\mathbf{b} K_{b\parallel} \nabla_{b\parallel})$

The expression for the parallel gradient above now enables the use of the support-operator method (SOM) [8,9,11,26,27] to obtain the parallel Laplacian, as found in Stegmeir et al. [21] in highly anisotropic diffusion. For any function $T \in \mathcal{H}^2$, and $\Psi \in \mathcal{H}^1$, both continuous in Ω, the Green formula reads:

$$\int_\Omega (\boldsymbol{\nabla} \cdot \mathbf{u}) \Psi dV + \int_\Omega \mathbf{u} \cdot \nabla \Psi dV = \int_\Gamma (\Psi \mathbf{u}) \cdot \mathbf{n} dS. \tag{12}$$

Considering $\mathbf{u} = -\mathcal{K} \nabla T$, the Green formula connects the gradient and the divergence operators. According to the definition of the $\mathcal{L}^2$-inner product in both scalar field and vector spaces H and $\mathbf{H}$, respectively, we write Eqs. (13), (14) for each one:

$$\langle -\nabla \cdot \mathcal{K} \nabla T, \Psi \rangle_H = -\int_\Omega \boldsymbol{\nabla} \cdot (\mathbf{u} \Psi) dV + \int_\Gamma (\Psi \mathbf{u}) \cdot \mathbf{n} dS. \tag{13}$$

$$\langle -\mathcal{K} \nabla T, \nabla \Psi \rangle_{\mathbf{H}} = \int_\Omega (\mathbf{u} \cdot \nabla \Psi) dV, \tag{14}$$

Eq. (12) establishes the connection between gradient and divergence operators as the self-adjoint one to each other. Considering any suitable boundary conditions (bi-periodic or homogeneous Dirichlet), the Green formula allows us to define the parallel diffusion operator directly from the parallel gradient as:

$$\langle -\nabla \cdot (\mathbf{b} \mathcal{K} \nabla_{b\parallel} T), \Psi \rangle = \langle \mathcal{K} \nabla_{b\parallel} T, \nabla_{b\parallel} \Psi \rangle, \tag{15}$$

Even if Eq. (15) is unambiguous at the continuous level, it involves two inner products, one defined in GS (Eq. (16)), and the other one in the FS (Eq. (17)) for any functions f and g as:

$$\langle f, g \rangle_{\mathrm{GS}} = \sum_\lambda f_\lambda \, g_\lambda \, \Delta V_\lambda, \tag{16}$$

$$\langle f, g \rangle_{\mathrm{FS}} = \sum_p \widetilde{f}_p \, \widetilde{g}_p \, \widetilde{\Delta V_p}. \tag{17}$$

According to Eq. (10), the inner product in FS can be estimated at the discrete level using evaluations of the diffusion on flux points denoted by $\{K_{b_\parallel p}\}$ as:

$$\langle [K] \nabla_{b\parallel} T, \nabla_{b\parallel} \Psi \rangle_{\mathrm{FS}} \approx \sum_p \widetilde{\Delta V_p} \left(K_{b_\parallel p} \sum_\lambda Q_{p\lambda} T_\lambda \right) \left(\sum_\mu Q_{p\mu} \Psi_\mu \right). \tag{18}$$

Depending on the number of contact surfaces a_p, a certain number of flux values can be associated for each λ. In terms of the SOM formalism [9], $Q_{p\lambda}$ of Eq. (8), defined in FS is here the *prime operator*. The discretization of the divergence (*derived operator* in terms of SOM) defined in FS into GS is the anti-adjoint of Q obtained by discrete analogue of Eq. (15). Then, the full operator $(\nabla \cdot [K] \nabla_{b\parallel})$ is endomorphic in GS. The left-hand side of Eq. (15) leads at the discrete level to:

$$\langle \nabla \cdot [K] \nabla_{b\parallel} T, \Psi \rangle \approx \sum_\sigma (\nabla \cdot [K] \nabla_{b\parallel} T)_\sigma \Psi_\sigma \Delta V_\sigma. \tag{19}$$

According now to Eqs. (15), (18) and (19), one deduces by identification that:

$$-(\nabla \cdot [K] \nabla_{b\parallel} T)_\lambda \approx \frac{1}{\Delta V_\lambda} \sum_p \left(K_{b_\parallel p} Q_{p\lambda} \sum_\mu Q_{p\mu} T_\mu \widetilde{\Delta V_p} \right). \tag{20}$$

The sum on μ term denotes the construction of the fluxes in FS from GS. Eq. (20) leads:

$$(\nabla \cdot [K] \nabla_{b\parallel} T) \approx -\Delta V^{-1} Q^T \widetilde{[K]} \widetilde{\Delta V_p} \, Q \, T \tag{21}$$

Upon multiplication by the cell volume ΔV, the SOM provides a symmetric discrete matrix: the product of an operator and its anti-adjoint is a self-adjoint, negative operator. The construction used by SOM reflects this symmetry and results in a symmetric matrix.

$$\mathcal{A}_{\lambda\mu} \Delta V_\lambda = \sum_p Q_{p\lambda} Q_{p\mu} [K]_p \widetilde{\Delta V_p} = \sum_p Q_{p\mu} Q_{p\lambda} [K]_p \widetilde{\Delta V_p} = \mathcal{A}_{\mu\lambda} \Delta V_\mu, \tag{22}$$

where $\widetilde{[K]_p} \widetilde{\Delta V_p} = \widetilde{[K \Delta V]}$ is a diagonal square matrix.

Please cite this article in press as: J.A. Soler et al., A new conservative finite-difference scheme for anisotropic elliptic problems in bounded domain, J. Comput. Phys. (2019), https://doi.org/10.1016/j.jcp.2019.109093

Finally, the conservativity of the scheme is verified by taking the special case $\Psi = 1$:

$$-\langle \nabla \cdot [K] \nabla_{b\parallel} T, 1 \rangle_{GS} = \langle [K] Q\, T, Q\, 1 \rangle_{FS} = 0$$

which is to say that the average of the parallel Laplacian over the computational domain is zero. This property follows from Eq. (11), which implies $Q \cdot 1 = 0$, and entails conservativity:

$$\langle (1 - \nabla \cdot [K] \nabla_{b\parallel}) T, 1 \rangle_{GS} = \langle T, 1 \rangle_{GS}$$

The proposed scheme therefore preserves three properties of the continuous operator, namely self-adjointness, positivity and conservativity [9,26].

3.4. Discretizations of the perpendicular gradient $\nabla_{b\perp}$ and Laplacian

In this paper, we also propose a conservative approach to discretize the operators in the perpendicular direction. In order to maintain the stencil size, the stencil that will be used matches that used in the parallel direction and defined in Sec. 3.2. The perpendicular gradient is estimated at the same points in FS, commonly shared with the surrounding control volumes (CVs). Indeed, perpendicular gradient is defined in FS from the discrete expression of its integral definition as follows:

$$\nabla_\perp T = \lim_{V(K) \to 0} \frac{1}{\|\mathbf{e}^1_\perp\| V(K)} \int_{S(K)} T \boldsymbol{e}^1_\perp \cdot \boldsymbol{n} dS \tag{23}$$

where $V(K)$ and $S(K)$ are the volume and the enclosing surface of the control volume K, which leads to the discrete definition:

$$\widetilde{\nabla_\perp T_p} = \frac{1}{\|\mathbf{b}^1_\perp\| \widetilde{\Delta V_p}} \sum_q T^{int}_{q\,p}\, a_{q\,p}\, \boldsymbol{b}^1_{\perp\, q\,p} \cdot \boldsymbol{n}_{q\,p}, \tag{24}$$

where $\widetilde{\Delta V_p}$ is the volume defined in Sec. 3.2, q the number of $\widetilde{\text{CV}}_p$ faces (see for example on Fig. 4), $\boldsymbol{b}_{\perp\, q\,p}$ the perpendicular vector to $\boldsymbol{b}$ (also perpendicular to the considered plane). Eq. (24) requires the evaluation of the flux $\widetilde{\nabla_{b\perp} T_p}$ at the barycenters of all faces. To maintain the stencil size, derivatives in y and z directions are evaluated at the points (x_i, y^-, z^-) and (x_{i+1}, y^+, z^-) in both $\mathcal{X}_i$ and $\mathcal{X}_{i+1}$ planes (Eq. (5)).

4. Construction of the stencils in a 2D domain

For simplicity, the construction is presented here in 2D but the generalization to 3D problems is straigthforward although cumbersome to write. The corresponding test case is presented in Sec. 5.3. The diffusion lines are also considered as straight and parallel, but the diffusion coefficient can be non uniform. The same notation as introduced in Sec. 3 are used, volumes and surfaces becoming surfaces and lengths, respectively. Since the aligned coordinates are local, the geometrical origin is established at T_{ij}.

4.1. Parallel Laplacian operator

4.1.1. Geometrical definitions

The definition of the fluxes space between adjacent control volumes is based on the geometrical definitions of the parallel diffusion lines $b_\parallel$. The local CV of $T_{i\,j}$ is bounded by the parallel field lines $b_{\parallel\, ij\pm 1/2}$ defined between $\mathcal{X}_{i\pm 1/2}$, Fig. 4. Considering the forward sense (+), the local CV is here in contact with two adjacent CV defined in the $\mathcal{X}_{i\pm 1}$ planes for the grid nodes $T_{i+1\, j+\xi}$ and $T_{i+1\, j+\xi+1}$ (note $\xi = 1$ in Fig. 4, see Eq. (46)). Considering $d_{b\parallel}$ the distance between two adjacent $\mathcal{X}$ planes when moving along the diffusion line as:

$$d_{b\parallel} = \frac{\Delta x}{\cos(\alpha)}, \tag{25}$$

the projection of $d_{b\parallel}$ in the y-direction writes:

$$y_{d_{b\parallel}} = d_{b\parallel} \sin(\alpha) = \Delta x \tan(\alpha).$$

The contact surfaces (Fig. 4b) lead to the following areas a_1 and a_2 (here lengths) such that:

$$a_1 = \Delta y + y_{d_{b\parallel}} - \Delta y\,(\xi) = d_{b\parallel} - (\xi)\Delta y$$

$$a_2 = \Delta y - a_1,$$

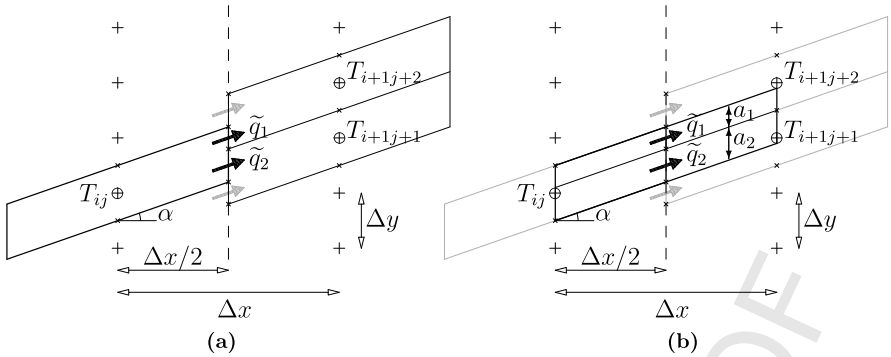


Fig. 4. (a) 2D chart used for the flux calculation at the barycenter of the contact surface between neighboring CVs. (b) Local $\widetilde{CV}$ defined by the diffusion lines surrounding the contact surfaces and bounded in the x-direction by the $\mathcal{X}_i$ and $\mathcal{X}_{i+1}$ planes.

The y-coordinates of the two barycenters of a_1 and a_2 express as:

$$y_{bc_1} = \frac{1}{2}(\Delta y + y_{d_{b\|}} - a_1)$$
$$y_{bc_2} = \frac{1}{2}(-\Delta y + y_{d_{b\|}} + a_2)$$

Then, the fluxes can be calculated at y_{bc_1} and y_{bc_2} in $x = (i + 1/2)\Delta x$. The control volume $\widetilde{CV}$ for each flux is limited by the parallel diffusion lines defined at $y =_{bc_1} \pm a_1/2$ and $y = y_{bc_2} \pm a_2/2$ (note $y_{bc_1} - a_1/2 = y_{bc_2} + a_2/2$, see Fig. 4).

4.1.2. Finite-differences step

The calculation of $\widetilde{q}_p$ reduces Eq. (8) to a finite-differences equation: the gradient of T at the barycenter is obtained by linear interpolation on b_{bc_p} at $\mathcal{X}_i$ and $\mathcal{X}_{i+1}$. The interpolation coordinates are obtained at x=$i\Delta x$ and $x = (i + 1)\Delta x$ as:

$$y^-_{int} = y_{bc_p} - \frac{d_{b\|}}{2},$$
$$y^+_{int} = y_{bc_p} + \frac{d_{b\|}}{2}.$$

Then, a linear interpolation in the y-direction allows us to evaluate T at $y^{\pm}_{int}$. For $p = 2$, it is illustrated on Fig. 5:

$$\begin{aligned} T^-_{int\,2} &= f^{int\,2}_{ij}\, T_{ij} + f^{int\,2}_{ij-1}\, T_{ij-1} \\ &= \left[1 - \frac{1}{\Delta y}\left|\left(y_{bc_2} - \frac{y_{d_{b\|}}}{2}\right)\right|\right] T_{ij} + \frac{1}{\Delta y}\left|\left(y_{bc_2} - \frac{y_{d_{b\|}}}{2}\right)\right| T_{ij-1} \end{aligned} \tag{26}$$

$$\begin{aligned} T^+_{int\,2} &= f^{int\,2}_{i+1j+\xi+1}\, T_{i+1j+\xi} + f^{int\,2}_{i+1j+\xi+1}\, T_{i+1j+\xi+1} \\ &= \left[1 - \frac{1}{\Delta y}\left(y_{bc_2} + \frac{y_{d_{b\|}}}{2} - \xi\Delta y\right)\right] T_{i+1j+\xi} + \frac{1}{\Delta y}\left(y_{bc_2} + \frac{y_{d_{b\|}}}{2} - \xi\Delta y\right) T_{i+1j+\xi+1} \end{aligned} \tag{27}$$

For $p = 1$ we get:

$$\begin{aligned} T^-_{int\,1} &= f^{int\,1}_{ij}\, T_{ij} + f^{int\,1}_{ij+1}\, T_{ij+1} \\ &= \left[1 - \frac{1}{\Delta y}\left(y_{bc_1} - \frac{y_{d_{b\|}}}{2}\right)\right] T_{ij} + \frac{1}{\Delta y}\left(y_{bc_1} - \frac{y_{d_{b\|}}}{2}\right) T_{ij+1} \end{aligned} \tag{28}$$

$$\begin{aligned} T^+_{int\,1} &= f^{int\,1}_{i+1j+\xi+1}\, T_{i+1j+\xi} + f^{int\,1}_{i+1j+\xi+1}\, T_{i+1j+\xi+1} \\ &= \left[1 - \frac{1}{\Delta y}\left(y_{bc_1} + \frac{y_{d_{b\|}}}{2} - \xi\Delta y\right)\right] T_{i+1j+\xi} + \frac{1}{\Delta y}\left(y_{bc_1} + \frac{y_{d_{b\|}}}{2} - \xi\Delta y\right) T_{i+1j+\xi+1}. \end{aligned} \tag{29}$$

Then, the gradient calculated at bc_p writes:

$$\nabla(T)_p = \frac{T^+_{int\,p} - T^-_{int\,p}}{d_{b\|}}.$$

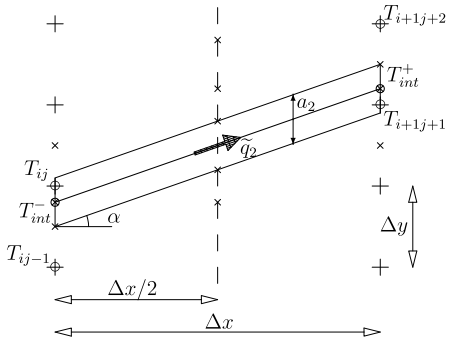


Fig. 5. 2-D chart of the calculation of the flux $\widetilde{q}_2$ attached to $\widetilde{CV}_2$. The volume $\widetilde{CV}_2$ is here a parallelogram with $\widetilde{\Delta V}_n = \Delta x a_n$. T^-_{int} and T^+_{int} are the values of the field linearly interpolated from T_{ij}, T_{ij-1} and $T_{i+1\,j+1}\,T_{i+1\,j+2}$ respectively (see Eqs. (26), (27), (28) and (29)).

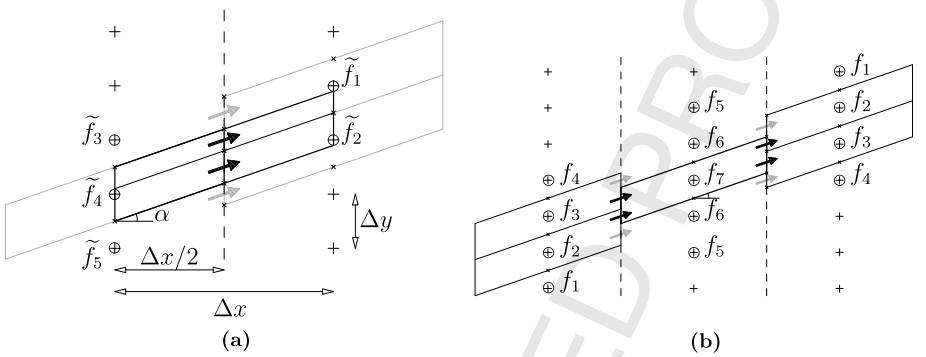


Fig. 6. Sketches of the stencils of the discrete gradient (a) and Laplacian (b) operators.

The value of $K_{b\|}$ at the barycenter is obtained here by interpolation in the same coordinates than $T^{\pm}_{int\,p}$ (Eqs. (26), (27), (28) and (29)), obtaining $K^-_{b\|\,int\,p}$ at $x = i\Delta x$ and $K^+_{b\|\,int\,p}$ at $x = (i+1)\Delta x$. The aligned interpolation leads to:

$$K_{b\|\,p} = \frac{K^-_{b\|\,int\,p} + K^+_{b\|\,int\,p}}{2} \tag{30}$$

Then, the discrete flux is obtained as:

$$\widetilde{q}_p = K_{b\|\,p} \nabla(T)_p \tag{31}$$

leading to the following stencil:

$$\begin{aligned}
\widetilde{f}_1 &= \frac{a_1}{a_1+a_2} K_{b\|\,1} f^{int\,1}_{i+1j+\xi+1} + \frac{a_2}{a_1+a_2} K_{b\|\,2} f^{int\,2}_{i+1j+\xi+1} \\
\widetilde{f}_2 &= \frac{a_1}{a_1+a_2} K_{b\|\,1} f^{int\,1}_{i+1j+\xi} + \frac{a_2}{a_1+a_2} K_{b\|\,2} f^{int\,2}_{i+1j+\xi} \\
\widetilde{f}_3 &= \frac{a_1}{a_1+a_2} K_{b\|\,1} f^{int\,1}_{ij+1} \\
\widetilde{f}_4 &= \frac{a_1}{a_1+a_2} K_{b\|\,1} f^{int\,1}_{ij} + \frac{a_2}{a_1+a_2} K_{b\|\,2} f^{int\,2}_{ij} \\
\widetilde{f}_5 &= \frac{a_2}{a_1+a_2} K_{b\|\,2} f^{int\,2}_{ij-1}
\end{aligned} \tag{32}$$

4.1.3. Complete stencil of the Laplacian operator

According to Eq. (21), the final stencil is given by the product of the transposed sparse matrix related to the flux definition on Eq. (31). The stencil with the coefficients of Eqs. (33) is shown on Fig. 6 . Due to the use of SOM, the resulting discrete Laplacian matrix is positive definite.

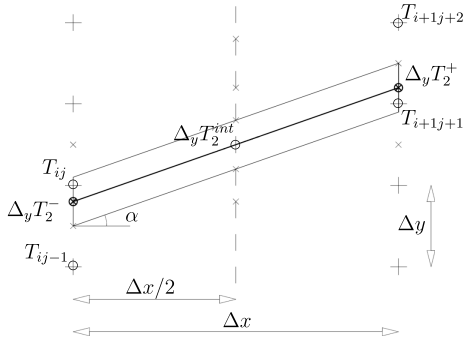


Fig. 7. Sketch including all quantities used to obtain the gradient in the y-direction in FS for $p = 2$ (Fig. 2). $\nabla_y T_2^-$ and $\nabla_y T_2^+$ are evaluated on $\mathcal{X}_i$ and $\mathcal{X}_{i+1}$, respectively, and further interpolated along the parallel direction to obtain $\Delta_y T_2^{int}$ at the barycenter.

$$
\begin{aligned}
f_1 &= -\widetilde{f}_1\widetilde{f}_5 \\
f_2 &= -\widetilde{f}_1\widetilde{f}_4 - \widetilde{f}_5\widetilde{f}_2 \\
f_3 &= -\widetilde{f}_4\widetilde{f}_2 - \widetilde{f}_1\widetilde{f}_3 \\
f_4 &= -\widetilde{f}_3\widetilde{f}_2 \\
f_5 &= -\widetilde{f}_3\widetilde{f}_5 \\
f_6 &= -\widetilde{f}_3\widetilde{f}_4 - f_5\widetilde{f}_4 - \widetilde{f}_1\widetilde{f}_2 \\
f_7 &= 1 - \widetilde{f}_4^2 - \widetilde{f}_5^2 - \widetilde{f}_1^2 - \widetilde{f}_2^2 - \widetilde{f}_3^2
\end{aligned}
\tag{33}
$$

The conservative discretization of the gradient (see an illustration on Fig. 4 together with SOM in the *present scheme* leads to a rather large stencil of 13-nodes. It can be compared to the 5-nodes stencil (with linear interpolation, or 7-nodes with an interpolation of degree 2) of the Ottaviani's scheme and to the 7-nodes stencil (with a linear interpolation) of the Stegmeir's scheme. Consequently, a slight overhead can be expected on the computational cost when inverting the matrix, although it is difficult to rigorously estimate it because depending on many other parameters like the mesh distribution, the pitch angle or the condition number (when using iterative solver).

4.2. Perpendicular gradient $\nabla_{b\perp}$ and Laplacian

Under the current assumptions, Eq. (24) reduces to:

$$
\widetilde{\nabla_{b\perp} T_p} = \frac{\widetilde{\nabla_y T_p} - \widetilde{\nabla_{b\parallel} T_p}\sin\alpha}{\cos\alpha},
\tag{34}
$$

where $\widetilde{\nabla_{b\parallel} T_p}$ is obtained using the aligned method described in Sec. 3.2. Considering for example $p = 2$, Fig. 7, the y-direction contribution is evaluated in the planes $\mathcal{X}_i$ and $\mathcal{X}_{i+1}$ as :

$$
\nabla_y T_2^- \approx K_{y2}^- \frac{T_{ij+1} - T_{ij}}{\Delta y},
\tag{35}
$$

and analogously for the gradient $\nabla_y T_2^+$, K_{y2}^- being the diffusion in the y-direction at bc_2 defined later. Finally, the gradient at the barycenter of $\widetilde{\mathrm{CV}}_p$ is obtained by interpolation along **b** as:

$$
\nabla_y T_2^{int} = \frac{1}{2}(\nabla_y T_2^- + \nabla_y T_2^+),
\tag{36}
$$

The value of the diffusion in the y-direction at bc_p is firstly evaluated at the stencil nodes, as (Fig. 7):

$$
K_{y\,ij} = K_{b\parallel}\sin(\alpha) + K_{b\perp}\cos(\alpha)
$$

Then, it is interpolated at bc_p using the stencil defined on Eq. (30) and based on the interpolation defined in Eqs. (26), (27), (28) and (29), Sec. 4.1.2. Once the perpendicular flux has been determined in FS, the full Laplacian is obtained by the SOM described in Sec. 3.3.

Please cite this article in press as: J.A. Soler et al., A new conservative finite-difference scheme for anisotropic elliptic problems in bounded domain, J. Comput. Phys. (2019), https://doi.org/10.1016/j.jcp.2019.109093

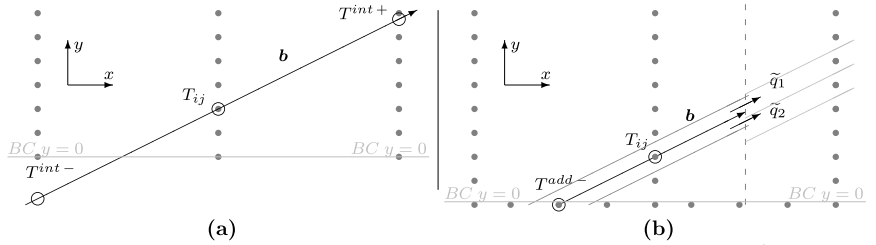


Fig. 8. Numerical discretizations near the boundary condition $y = 0$ using *aligned methods*. (a) *Aligned method* with ghost points T^{int-} located along the parallel direction and possibly far outside the domain (extrapolated points method). (b) Present method with ghost points T^{add} added on the boundary of the domain.

4.3. Numerical discretization of the boundaries in 2D

We consider here bounded domains where the flow intercepts the boundaries in the direction of anisotropy (parallel direction). The discretization presented above must be adapted to keep the accuracy while remaining compatible with the discretization adopted for inner nodes. As a reminder, aligned methods allow the reduction of required degrees of freedom in one mesh direction, by using the knowledge that the main contributions to the solutions will either be uniform or slowly varying in the main diffusion direction. One can then use a mesh with fine resolution in one direction, so as to resolve the potentially fast variations in the perpendicular direction, but coarse resolution in the other direction that accounts for variations in the parallel direction. The constraint to be taken into account in order to maintain these advantages is that the mesh on the boundary allows to adequately represent the fast variations of the solution along the boundary itself. The *non-aligned methods* [7], or the *aligned methods* using a stencil based on surrounding grid points [4], keep working near the boundaries, possibly using few ghost points (since the stencil are not oriented to $\boldsymbol{b}$). However, near the boundary, *aligned methods* for which the stencil is oriented, such as the present scheme, or in others Refs. [18,19,21], parallel diffusion line tracking intercepts the neighbor plane outside the domain limits, Fig. 8a. This reason suggests another treatment of the operator near the boundary, solving Eq. (1) with an aligned approach, avoiding the uncertainties of far ghost points.

We propose here to add ghost points directly on the boundary of the domain, and located at the intersection of the parallel diffusion lines with the boundary, as shown on Fig. 8b. Such points are needed since, depending on the resolution and the incidence α, we may have at the two boundaries:

$$\int_{x_i}^{x_{i-1}} \frac{b_y}{b_x} dx < 0 \quad \text{at} \quad y = 0, \qquad \text{and} \qquad \int_{x_i}^{x_{i+1}} \frac{b_y}{b_x} dx > 2\pi \quad \text{at} \quad y = 2\pi\,.$$

For any point T_{ij} located close to the grid limits at $y = 0$ and $y = 2\pi$, extra points are added in the x-direction at the coordinates:

$$x^{add}_{(y=0)} = i\Delta x - \frac{j\Delta y}{\tan(\alpha)} \qquad \text{and} \qquad x^{add}_{(y=2\pi)} = i\Delta x + \frac{(N_y - j)\Delta y}{\tan(\alpha)}. \tag{37}$$

These points being now located on the boundary, the value of the field may be directly obtained from the boundary conditions associated to Eq. (1). For Dirichlet boundary condition ($\beta = 0$ and $\gamma = 1$) the result is immediate. For Neumann boundary condition ($\beta = 1$ and $\gamma = 0$), the values of $T^{add-}_{(y=0)}$ and $T^{add+}_{(y=2\pi)}$ are here obtained by the derivative in the parallel direction evaluated using interior grid points at $y = 0$ and $y = 2\pi$, respectively. In this case, we get:

$$\nabla_{b\|} T^{add-}_{(y=0)} = \frac{(d^2_{b\|\,T} + d^2_{b\|\,1})T^{add-}_{(y=0)} - d^2_{b\|\,T} T_{ij} + d^2_{b\|\,1} T^{int+}}{d^2_{b\|\,1} d_{b\|\,T} - d_{b\|\,1} d^2_{b\|\,T}}, \tag{38}$$

$$\nabla_{b\|} T^{add+}_{(y=2\pi)} = \frac{(d^2_{b\|\,2} - d^2_{b\|\,T})T^{add+}_{(y=2\pi)} + d^2_{b\|\,T} T_{ij} - d^2_{b\|\,2} T^{int-}}{d^2_{b\|\,2} d_{b\|\,T} - d_{b\|\,2} d^2_{b\|\,T}}, \tag{39}$$

where $d_{b\|\,1}$ and $d_{b\|\,2}$ are the lengths in the parallel direction between T^{add-} and T_{ij} at $y = 0$, and between T_{ij} and T^{add+} at $y = 2\pi$ respectively:

$$d_{b\|\,1} = \frac{j\Delta y}{\sin(\alpha)} \qquad \text{and} \qquad d_{b\|\,2} = \frac{(N_y - j)\Delta y}{\sin(\alpha)}, \tag{40}$$

with $d_{b\|\,T} = d_{b\|\,1} + d_{b\|}$ in Eq. (38) and $d_{b\|\,T} = d_{b\|\,2} + (1/2)d_{b\|}$ in Eq. (39). Let's notice that in this case $d_{b\|\,p}$ is the simplification of $a_p/\Delta V_p$ (Eq. (8)). $T^{int\pm}$ defines the value of the field interpolated in the plane $\mathcal{X}_{i\pm1}$.

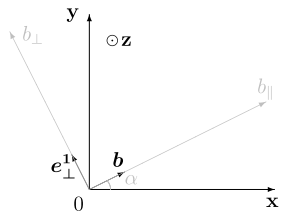


Fig. 9. Directions of the principal axes of the diffusion tensor in the (x, y)-plane. α defines the misalignment angle of the principal axes with respect to grid points directions.

Since the values of T_{ij}^{add} at $y = 0$ and $y = 2\pi$ are located along $\boldsymbol{b}$, the CV associated to T_{ij} (See on Fig. 1) is aligned with the control volume associated to T_{ij}^{add}. The flux discretized using finite-differences between T_{ij} and T_{ij}^{add} remains conservative. The complete operator can be calculated by considering the fluxes balance in the CV of T_{ij}, i.e.:

$$\nabla \cdot [K] \nabla_{b\parallel} T_{ij} = \frac{\frac{1}{a_T}(a_1 \widetilde{q}_1^{+} + a_2 \widetilde{q}_2^{+}) - q^{add-}}{\frac{1}{2}\left(d_{b\parallel 1} + d_{b\parallel}\right)}, \tag{41}$$

$$\nabla \cdot [K] \nabla_{b\parallel} T_{ij} = \frac{q^{add+} - \frac{1}{a_T}(a_1 \widetilde{q}_1^{-} + a_2 \widetilde{q}_2^{-})}{\frac{1}{2}\left(d_{b\parallel} + d_{b\parallel 2}\right)}, \tag{42}$$

where $q_p^{\pm}$ is the total flux considering the fluxes obtained in Eq. (31), and q_{ij}^{add-} and q_{ij}^{add+} are the fluxes between the inner grid point (i, j) and the corresponding grid point added on the boundary:

$$q^{add-} = K_{b\parallel}^{add-} \frac{T_{ij} - T_{(y=0)}^{add-}}{d_{b\parallel 1}}, \quad \text{and} \quad q^{add+} = K_{b\parallel}^{add+} \frac{T_{(y=2\pi)}^{add+} - T_{ij}}{d_{b\parallel 2}}, \tag{43}$$

where the values of the parallel diffusion are obtained by linear interpolation such that:

$$K_{b\parallel}^{add-} = \frac{K_{b\parallel\,(y=0)}^{add-} + K_{b\parallel\,ij}}{2} \quad \text{and} \quad K_{b\parallel}^{add+} = \frac{K_{b\parallel\,ij} + K_{b\parallel\,(y=2\pi)}^{add+}}{2}.$$

Note that the discretization of the Laplacian here does not use SOM. Then the positive-definite and self-adjoint properties are not proven in this case.

5. Numerical tests

Numerical tests have been performed on Eq. (1) in 2D in the (x, y)-plane. Thus, the rotation matrix $\boldsymbol{\mathcal{R}}$ of Eq. (2) defines as:

$$\boldsymbol{\mathcal{R}} = \begin{bmatrix} \cos\alpha & -\sin\alpha \\ \sin\alpha & \cos\alpha \end{bmatrix},$$

where the angle α measures the non-alignment between the unit vector along the anisotropy direction $\boldsymbol{b}$ and the x-axis. Thus, it results $\boldsymbol{b} = (\cos\alpha, \sin\alpha, 0)^t$ (Fig. 9).

Let's notice that α may vary in x in some tests leading to a non constant $\boldsymbol{b}$.

In all tests $\mu = 1$ (see a discussion in Appendix A), except for the special case of the Poisson's equation.

The following discretization methods have been considered for the tests:

- The *classical method*, referring to the asymmetric approach [4].
- The *Günter's method*, referring to the symmetric approach proposed by Günter et al. [7].
- The *Ottaviani's method*, referring to an aligned approach (oriented stencil) based on a second-order parallel, polynomial interpolation [18,19].
- The *Stegmeir's method*, referring to an aligned approach based on a linear interpolation [21].
- The *present method* referring to the work done in this paper. It extends *Stegmeir's* method to a conservative discretization in both parallel and perpendicular directions and to an efficient discretization of the boundary condition in bounded domains.

The first two methods used stencils independent of the diffusion tensor, and thus lie in the class of *non-aligned methods*. In contrast, the other methods lie in the class of *aligned methods*, as defined in Sec. 2.

When involved in the tests, the perpendicular part of the diffusion operator is discretized using the scheme proposed in this work in Sec. 3.4 for both *Stegmeir's method* and *Ottaviani's method*, which originally do not address the discretization of this flow direction.

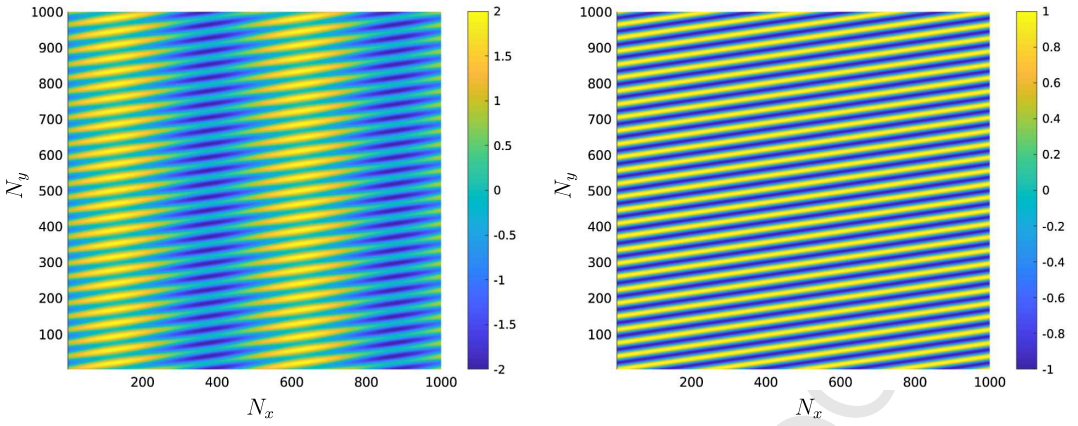


Fig. 10. 2D plots of T_a when solving Eq. (1) analytically with source term S_a Eq. (44) for straight parallel magnetic field lines ($\zeta = 0$): (a) an isotropic case $K_{b_\parallel} = K_{b_\perp}$ and (b) a strongly anisotropic case $K_{b_\parallel} = 10^6 K_{b_\perp}$. The source term S_a has following parameter values: $C_1 = 0$, $C_2 = 1$, $C_3 = 1$, $m_y = 27$, $m_{x,1} = 4$ and $m_{x,2} = 2$. (For interpretation of the colors in the figure(s), the reader is referred to the web version of this article.)

5.1. Numerical details

The following manufactured source term S_a has been considered, corresponding to the superposition of a constant, an aligned and a non-aligned contribution:

$$S_a(x, y) = C_1 + C_2 \cos\left[m_y y + (m_{x,1} x + \zeta \sin x)\right] + C_3 \sin\left(m_{x,2} x\right) \tag{44}$$

This source term corresponds indeed to the superposition of fluctuations varying rapidly in the perpendicular direction while being uniform along the parallel direction. The angle $\phi = tan^{-1}((m_{x,1} + \zeta \cos x)/m_y)$ defines the orientation of the aligned modes. Let's notice it may vary with x when $\zeta \neq 0$ that corresponds to curved magnetic field lines defined by $\mathbf{b}(x, y) = (-m_y, m_{x,1} + \zeta \cos x)$. The parameter ζ quantifies the magnitude of the curvature. The non-parallel modes vary only in x. In the case where $\phi = \alpha$, α being the pitch angle, the resolution of Eq. (1) with $K_{b\perp} = 0$ leads to the following solution:

$$T_a(x, y) = C_1 + C_2 \cos\left(m_y y + (m_{x,1} x + \zeta \sin x)\right) + \frac{1}{1 + K_{b\parallel} m_{x,2}^2} C_3 \sin\left(m_{x,2} x\right) \tag{45}$$

The fluctuations related to the first term should then dominate, and the damping coefficient of this particular contribution is a good indicator of the quality of the discretization used.

Tests are made fixing resolutions in the x-direction ($N_x = 8$, 16 and 32) while varying N_y such that $(N_{dof})_{max} = \max(N_x \cdot N_y) = 512^2$.

Most of the comparative tests with former works of the literature have been performed for a constant pitch angle corresponding to $\zeta = 0$ in Eq. (44), and thus straight parallel magnetic field lines in Sec.5.3: $\alpha = \tan^{-1}(4/27)$, $m_y = 27$, $m_{x,1} = 4$, $m_{x,2} = 2$, $C_1 = 0$, $C_2 = 1$, $C_3 = 0.25$. 2D plots of T_a in this configuration are shown on Fig. 10 for both a small ($K_{b_\parallel} = K_{b_\perp}$) and a large ($K_{b_\parallel} = 10^6 K_{b_\perp}$) anisotropy, showing or not parallel modulations in the parallel direction, respectively.

However, in order to show the capability of the present method to deal also with non uniform $\mathbf{b}$, some accuracy tests have been performed for curved magnetic field lines in Sec. 5.5 for $\zeta \neq 0$. 2D plots of T_a in this configuration are shown on Fig. 11 for $K_{b_\parallel} = 10^6$, $K_{b_\perp} = 0$, and for two magnitudes of the curvature $\zeta = 4$ and 8.

5.2. Error estimate

Numerical tests in Appendix B show that the $\mathcal{H}^1$-error (Appendix C) is better suited than the classically used $\mathcal{L}^2$-error to evaluate the accuracy of these schemes. Indeed, the $\mathcal{L}^2$-error eventually leads to misleading behavior due to some eventual spurious aliasing effect.

Tests are made fixing resolutions in the x-direction while varying N_y. For all tests, only the optimal values of the error are retained.This is illustrated on Fig. 12a for straight parallel magnetic field lines corresponding to $\zeta = 0$. They correspond to the minimal N_{dof} and H^1-error relation. For each fixed resolution in the x-direction, the error is foremost dominated by the interpolation error in the y-direction for the discretization of the parallel Laplacian of the aligned fluctuations, and decreases when N_y increases. Let's notice that for a given value of N_{dof} the smallest error is obtained with the smallest resolution in the x-direction since it is associated to the largest resolution in the y-direction.

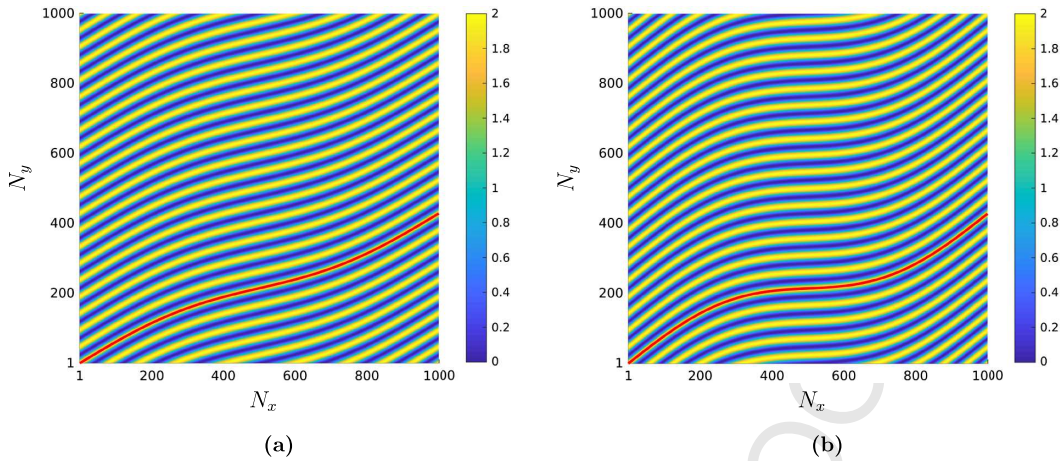


Fig. 11. 2D plots of T_a when solving Eq. (1) analytically with source term S_a Eq. (44) for curved magnetic field lines ($\zeta \neq 0$) and $K_{b_\parallel} = 10^6$, $K_{b_\perp} = 0$: (a) $\zeta = 4$ and (b) $\zeta = 8$. The red curves show two magnetic field lines. The source term S_a has following parameter values: $C_1 = 0$, $C_2 = 1$, $C_3 = 0$, $m_y = 21$, $m_{x,1} = 9$.

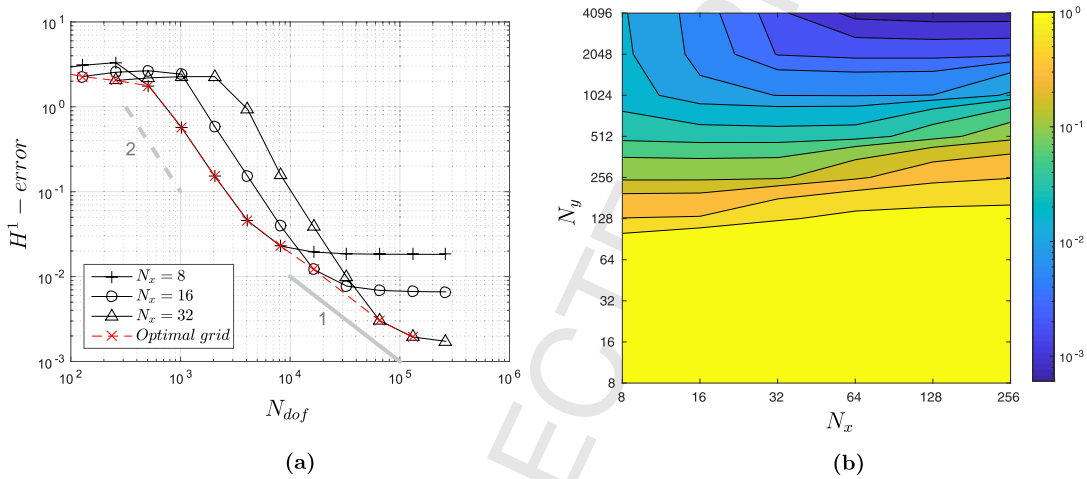


Fig. 12. Error estimate. (a) $\mathcal{H}^1$-error convergence for the *present method* when increasing N_{dof} for three resolutions in the x-direction. For each value of the error, only the point corresponding to the lowest resolution is kept (red dotted line). (b) 2D plot of the $\mathcal{H}^1$-error as a function of N_x and N_y. Bi-periodic domain with $K_{b_\parallel} = 10^0$. Straight parallel magnetic field lines, $\zeta = 0$.

From a certain value of N_{dof} (which depends on N_x), the error stops decreasing, and becomes dominated by the error made in discretizing the parallel Laplacian of non-aligned fluctuations of the solution. N_x being fixed implies that the parallel step-size is constant, and the error therefore converges to a constant value.

Fig. 12b presents the error as a function of N_x and N_y. It highlights the property of aligned methods which is that the accuracy of the solution depends only weakly on the parallel step-size, and hence on N_x, for moderate resolutions. A much larger improvement in accuracy can be gained by increasing N_y, Fig. 12b indicates that increasing N_x (parallel resolution) only improves the accuracy for sufficient resolutions in the y-direction.

5.3. Accuracy tests in a 2D periodic domain and for straight parallel magnetic field lines

Eq. (1) with $\mu = 1$ and periodic boundary conditions in x and y direction is considered. For these tests $\zeta = 0$.

5.3.1. Accuracy tests for a non-zero $K_{b_\perp}$

Convergence results are presented on Fig. 13 for an isotropic ($K_{b_\parallel} = K_{b_\perp}$) and an anisotropic ($K_{b_\parallel} = 10^6 K_{b_\perp}$) diffusion tensor (Eq. (2)). When the tensor is isotropic, there is no significant difference between all the methods, and the errors converge at nearly the same rate, Fig. 13a. However, when the tensor becomes anisotropic, Fig. 13b clearly shows the superiority of the *aligned methods*, owing to their better ability to capture the uniformity of the dominant contribution along the vector **b**. As expected by construction, the *present method* behaves similarly in this case as the two other *aligned*

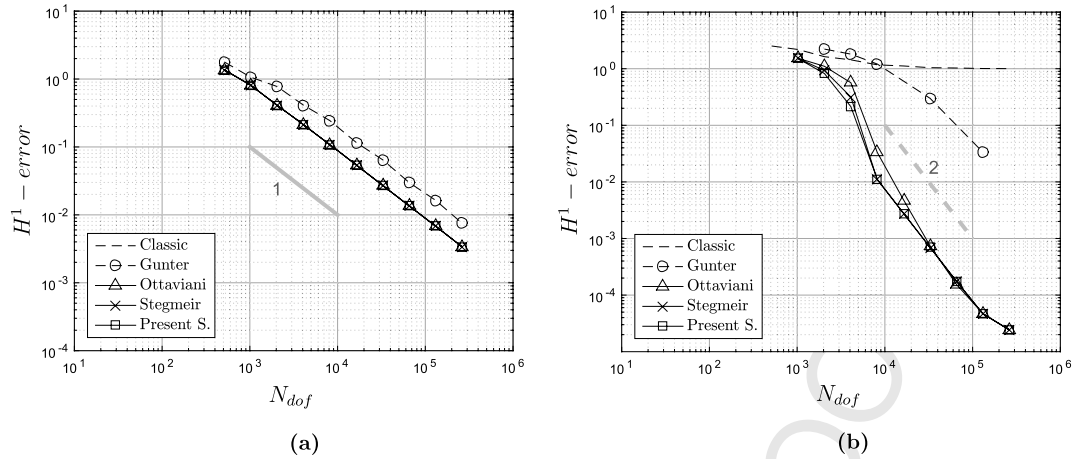


Fig. 13. $\mathcal{H}^1$-error convergence for an isotropic ($K_{b_\parallel} = K_{b_\perp}$) (a) and an anisotropic ($K_{b_\parallel} = 10^6 K_{b_\perp}$) (b) diffusion tensor. Bi-periodic computational domain.

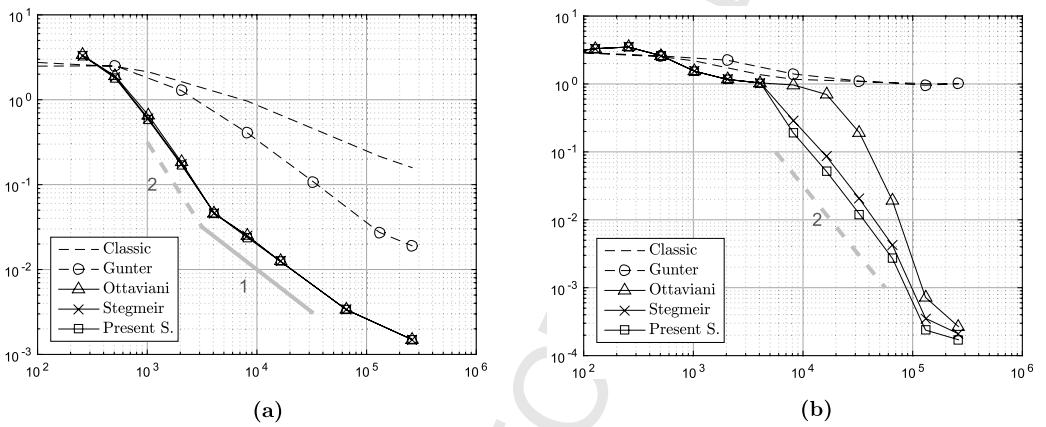


Fig. 14. $\mathcal{H}^1$-error convergence in a bi-periodic computational domain with $K_{b_\perp} = 0$ and (a) $K_{b_\parallel} = 10^0$ and (b) $K_{b_\parallel} = 10^6$.

methods of Ottaviani and Stegmeir. However, the *classical method* fails to converge, and though it converges, *Günter's method* requires many more points for a given accuracy.

5.3.2. Accuracy tests for $K_{b_\perp} = 0$

We now focus on the parallel flux estimate, which is the largest source of error in such computation, and we assume that $K_{b_\perp} = 0$. Convergence results are presented on Fig. 14 for two values of the parallel diffusion, $K_{b_\parallel} = 10^0$ and 10^6. These values correspond to solutions where parallel fluctuations are weakly or strongly damped, respectively.

For $K_{b_\parallel} = 10^0$ (Fig. 14a), the *present method* behaves as the other two *aligned methods* of the literature, and leads to a much better convergence rate than *non-aligned methods*. In addition, the three *aligned methods* need fewer points for a given accuracy, illustrating their superior ability to accurately compute the parallel Laplacian. For an error of about 10^{-2}, they indeed need about 10 times fewer points. The shift in the convergence rate between 2 and 1 at $N_{dof} = 4096$, already observed on Fig. 12, corresponds to change in the structure of the error. For $N_{dof} \leqslant 4096$, the error is dominated by the interpolation error in the y-direction, required by all *aligned methods* to evaluate the parallel gradient. This error decreases when N_y increases, the resolution in the x-direction being fixed. For $N_{dof} > 4096$, the error does no longer depend on the resolution in the y-direction but only on the resolution in the x-direction.

For $K_{b_\parallel} = 10^6$ (Fig. 14b), *non-aligned methods* fail to converge regardless of the resolution. In this case, the *aligned methods*' trend is fully related with the interpolation method, the fluctuations with parallel variations being strongly damped (third component of Eq. (44) vanishes), the problem becomes a fully aligned problem constant in the parallel direction. The *present method* provides the best result on this test. The difference with *Stegmeir's method* is small whatever the resolution, but it is larger with *Ottaviani's method*, particularly at moderate resolutions. The difference with this latter decreases at high resolutions.

Please cite this article in press as: J.A. Soler et al., A new conservative finite-difference scheme for anisotropic elliptic problems in bounded domain, J. Comput. Phys. (2019), https://doi.org/10.1016/j.jcp.2019.109093

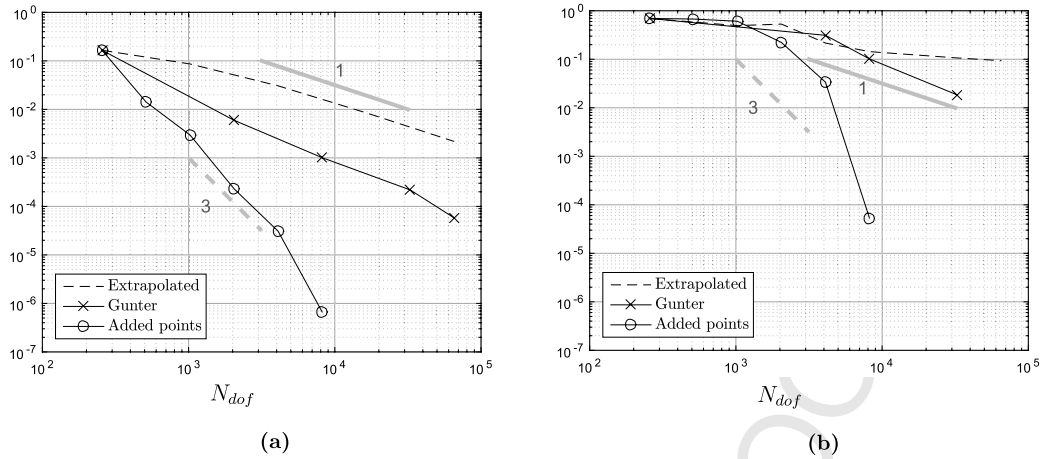


Fig. 15. $\mathcal{L}^2$-error convergence of the *present method* for $K_{b_\parallel} = 1$ (a) and $K_{b_\parallel} = 10^6$ (b) in a bounded domain and for three discretizations of the Dirichlet boundary condition. $K_{b_\perp} = 0$.

5.4. Accuracy tests in a 2D bounded domain and for straight parallel magnetic field lines

Another new feature presented here is an efficient and accurate discretization of the boundary conditions in bounded domains for aligned methods, which are involved in many realistic applications although they have been much less investigated in the relevant literature.

Eq. (1) with $\mu = 1$ and Dirichlet boundary conditions is considered. In all tests, the perpendicular diffusion $K_{b_\perp} = 0$.

Three discretizations of the boundary condition have been formerly mentioned in Sec. 4.3: added points aligned along b (extrapolated grid points in the y-direction, Fig. 15a), *Günter method* which relies solely on points already in the grid, and finally added points on the boundary (added points), which is the new discretization proposed in this paper (Fig. 15b).

On Fig. 15 the *present method* is tested with the three discretizations of the boundary condition. The $\mathcal{L}^2$-error has been retained as the blending of aligned methods for interior points and non-aligned methods in the neighborhood of the boundaries, which use different discretizations of gradients, makes the evaluation of the $\mathcal{H}^1$-error problematic. Furthermore, the $\mathcal{L}^2$-error is sufficient here to qualify the differences in accuracy between the proposed boundary discretizations. Extrapolated grid points in the y-direction shows poor performances, in particular for $K_{b_\parallel} = 10^6$, since the rapid variations in the parallel direction limit the accuracy of the extrapolation, as explained in Sec. 4.3. The use of a non-aligned approach like with *Günter's method* in the discretization of the boundary condition needs only one ghost point in the y-direction, but the ratio N_x/N_y is out of the limit of Nyquist-Shannon theorem provided for *non-aligned methods* when *aligned methods* reach higher performance. The added points discretization proposed in this work maintains the convergence found in bi-periodic cases even if it slightly increases the number of global unknowns required. Indeed, the number of added points is equal to $2N_x\xi(\ll N_x \times N_y)$, where ξ defines the shift of the grid as:

$$\xi = \lfloor \frac{\Delta x}{\Delta y} \tan\alpha \rfloor, \tag{46}$$

considering **b** and the diffusion tensor [K] as uniform in Ω.

On Fig. 16, accuracy tests are presented for all methods. The added points discretization proposed here is used for the three *aligned methods* (Ottaviani, Stegmeir and present). This new discretization of the boundary works well whichever the *aligned method* used in the interior of the domain. It allows to recover the good general trends obtained for the bi-periodic configuration (Fig. 14). As previously, the *present method* associated to this new discretization of the boundary provides the best results.

5.5. Accuracy tests in a 2D periodic domain and for curved magnetic field lines

Eq. (1) with $\mu = 1$ and periodic boundary conditions in x and y directions is considered here with a non uniform magnetic field **b** such that:

$$\boldsymbol{b}(x, y) = \begin{pmatrix} -m_y \\ m_{x,1} + \zeta \cos x \end{pmatrix} \tag{47}$$

Two tests have been performed for $\zeta = 4.$ and $\zeta = 8$. The following parameter values are used: $C_1 = 0$, $C_2 = 1$, $C_3 = 0$, $m_y = 21$, $m_{x,1} = 9$.

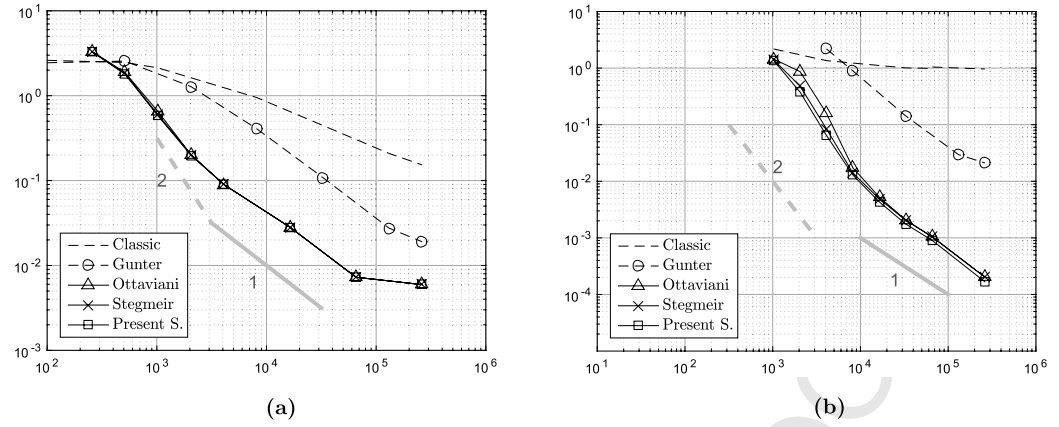


Fig. 16. $\mathcal{H}^1$-error convergence in a bounded domain with Dirichlet boundary conditions for all methods. For all *aligned methods* the added points discretization is used. (a) $K_{b_\parallel} = 10^0$ and (b) $K_{b_\parallel} = 10^6$.

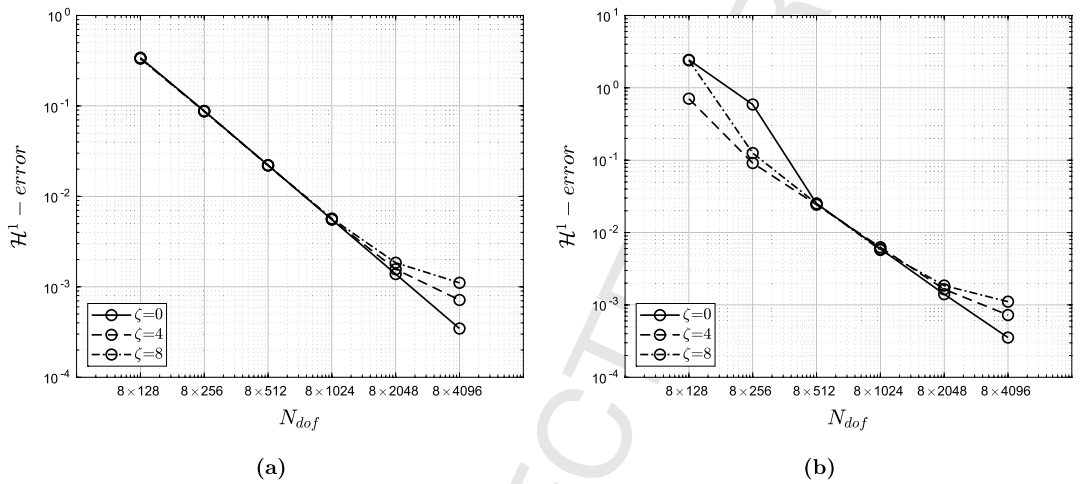


Fig. 17. $\mathcal{H}^1$-error convergence in a bi-periodic domain for the *present method* and for different ζ-values. **(a)** $K_{b_\parallel} = 1$ and **(b)** $K_{b_\parallel} = 10^6$.

Fig. 17 shows the convergence of the error in the $\mathcal{H}^1$-norm for the *present method*. Results are presented for both straight parallel and curved magnetic field lines. The results show the good behavior of the *present method* to deal with non uniform magnetic field, since the convergence is only very slightly affected by the curvature when varying N_y. This feature of the method is very encouraging for a future implementation in a code simulating tokamak plasmas.

5.6. Tests of conservativity

A new feature of the *present method* with respect to existing *aligned methods* of the literature is to involve a conservative discretization of the fluxes. It is shown here for the parallel operator, the discretization of the perpendicular operator (Sec. 3.4) implicitly guaranteeing the conservativity in this direction.

In *aligned methods* in the literature, the *Ottaviani's* and *Stegmeir's methods* evaluate fluxes at the center of the CVs faces for each plane $\mathcal{X}_i$ leading to a misalignment of the fluxes between adjacent CVs (Fig. 18a). On the contrary, the discretization of the fluxes calculated at the center of the common faces of two adjacent CVs ensures here the conservativity of the *present method* within the domain, Fig. 18b. This flux definition leads to a symmetric definition of fluxes between two adjacent $\mathcal{X}$-planes, independently of $K_{b\parallel}$. To show that, a test has been carried out in a 2D periodic domain considering $T_a = 2 + \sin(x)\sin(y)$, with a non homogeneous parallel diffusion, $K_{b\parallel} = 2 + \sin(x)\sin(y)$. Varying the number of degrees of freedom, N_{dof}, we plot the quantity $|\widetilde{q}_i^+| - |\widetilde{q}_{i+1}^-|$ that represents the balance between the red and blue fluxes of Fig. 18. The Fig. 18c shows that only the *present method* leads to a perfect zero balance of the fluxes inside the domain whatever the resolution. For the two other methods, there is always a small unbalanced between the fluxes that rises at low resolutions.

This property can be actually extended to any bounded domain. Indeed, the addition of extra points aligned with the inner node along the parallel direction (Sec. 4.3) also generates a conservative definition of the fluxes across the boundary. This is shown here by considering a y-bounded domain with Dirichlet boundary conditions. The pitch angle and the grid

Please cite this article in press as: J.A. Soler et al., A new conservative finite-difference scheme for anisotropic elliptic problems in bounded domain, J. Comput. Phys. (2019), https://doi.org/10.1016/j.jcp.2019.109093

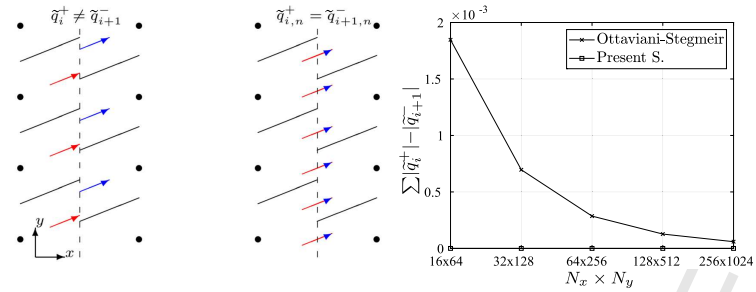


Fig. 18. Sketches showing flux discretizations between adjacent control volumes for *Ottaviani's method* (a) aligned to the grid points [18] and (b) the *present method* centered to the contact surface between control volumes. (c) Plot for different grids $N_x \times N_y$ of the relative difference $\sum |\tilde{q}_i^+| - |\tilde{q}_{i+1}^-|$ between the forward fluxes of $\mathcal{X}_i$ plane (red fluxes in sketch (a)) and the backward fluxes of $\mathcal{X}_{i+1}$ plane (blue fluxes in sketch (a)).

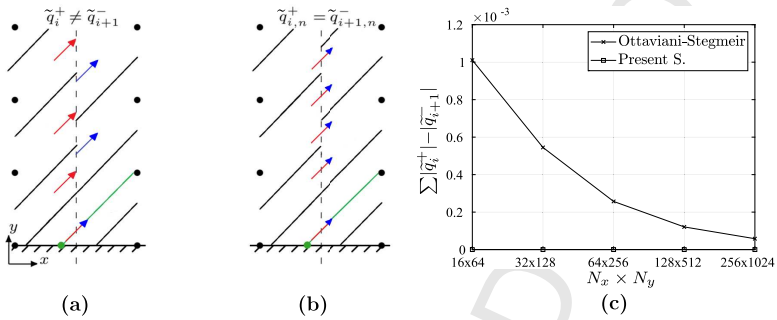


Fig. 19. Sketches showing flux discretizations between adjacent control volumes in bounded domain for *Ottaviani's method* aligned to the grid points [18] (a) and the *present method* centered to the contact surface between control volumes (b). The added point (green) on the boundary is aligned along the parallel direction to the grid inner point. (c) Plot for different grids $N_x \times N_y$ of the relative difference $\sum |\tilde{q}_i^+| - |\tilde{q}_{i+1}^-|$ between the forward fluxes of $\mathcal{X}_i$ plane (red fluxes on (a, b)) and the backward fluxes of $\mathcal{X}_{i+1}$ plane (blue fluxes on (a, b)) considering Dirichlet boundary conditions.

nodes relationship ($N_y = 4N_x$) enforce a $shift = 1$, leading to extra points added at the boundary and aligned with $N_y = 2$ and $N_y = N_{y,max} - 1$ grid nodes (see in Sec. 4.3). The zero fluxes balance shown on Fig. 18 confirms this feature of the *present method*. Calculating the flux balance up to the boundary even reduces the unbalance of the Ottaviani's and Stegmeir's method with respect to the bi-periodic case.

5.7. A case of prime importance: the Poisson equation

A special case of prime importance in many physical models (for example in the search of stationary solutions to the heat equation) is the case of the Poisson's equation. We consider here the general case where it is associated to Robin boundary conditions such that:

$$-\nabla \cdot (\mathcal{K} \cdot \nabla))T = S \qquad \text{on } \Omega \tag{48}$$

$$\frac{1}{R}\nabla_{\parallel} T + T = s \qquad \text{in } \Gamma \tag{49}$$

All values of R lead to regular problems. However, if the limit $R \to +\infty$ is very well-behaved, as one then approaches a problem with Dirichlet boundary conditions, the limit $R \to 0$ can be demanding, as one then approaches a problem with Neumann boundary conditions which is known to be-ill posed. Robin boundary conditions with $R = 1$ and $R = 10^{-3}$ are tested here. See Fig. 19.

For $R = 1$, the weight of the Dirichlet and the Neuman part in the Robin boundary condition is the same. Fig. 20 shows that *aligned methods* associated to the added points approach proposed in this paper for the discretization of the boundary condition confirm their superiority for both low ($K_{b_{\parallel}} = 10^0$) and large ($K_{b_{\parallel}} = 10^6$) parallel diffusion. As previously, the *present method* tends to provide the best results, even if the differences with the *Stegmeir's* and *Ottaviani's method* are small in this case. For both values of the diffusion, the *classical method* does not converge. The *Günter's method* tends to behave slightly better than for the tests carried out in the periodic domain, and the *classical* and *Günter's methods* seem to give errors independent of the level of anisotropy.

For $R = 10^{-3}$, the Neumann part becomes dominant over the Dirichlet part in the Robin boundary conditions. As mentioned above, the resolution of the Poisson's equation becomes much more demanding. This appears on the results shown on Fig. 21. The *classical method* does not converge whatever the parallel diffusion (as for the case $R = 1$), and, if the *Günter's*

Please cite this article in press as: J.A. Soler et al., A new conservative finite-difference scheme for anisotropic elliptic problems in bounded domain, J. Comput. Phys. (2019), https://doi.org/10.1016/j.jcp.2019.109093

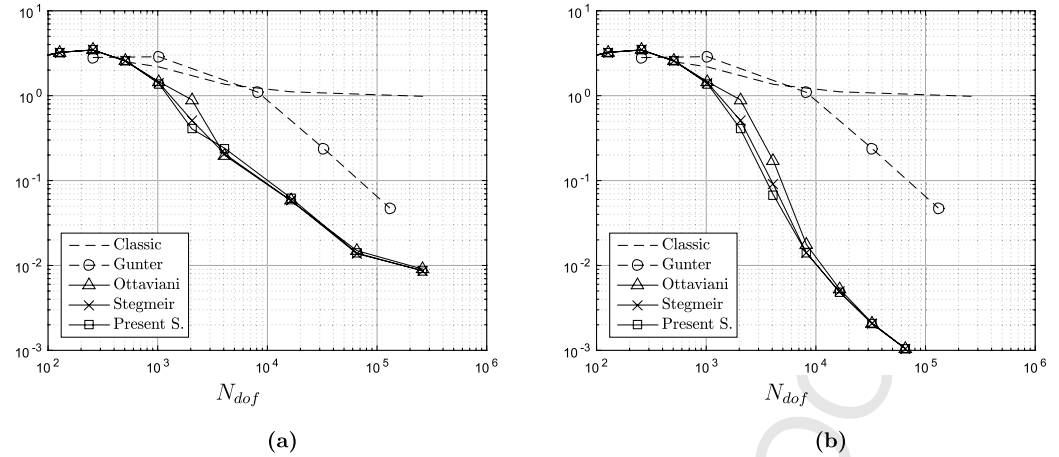


Fig. 20. $\mathcal{H}^1$-error convergence for the Poisson's equation Eq. (48) with Robin boundary condition with $R = 1$. $K_{b_\parallel} = 10^0$ (a) and $K_{b_\parallel} = 10^6$ (b). $K_{b_\perp} = 0$.

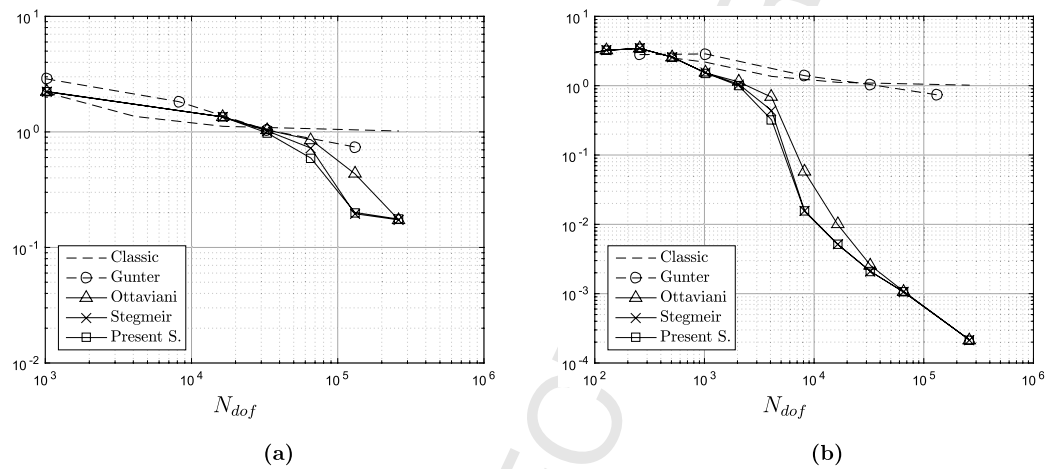


Fig. 21. $\mathcal{H}^1$-error convergence for the Poisson's equation Eq. (48) with Robin boundary condition with $R = 10^{-3}$. $K_{b_\parallel} = 10^0$ (a) and $K_{b_\parallel} = 10^6$ (b). $K_{b_\perp} = 0$.

method continues to converge, its convergence rate is strongly reduced. The *aligned methods* visibly still fare better than the *non-aligned methods*. As expected, increasing the anisotropy of the diffusion tensor improves their efficiency. The *present method* provides here similar results to the *Stegmeir's method*.

6. Conclusion

In this work, we have proposed a new finite-difference scheme to solve highly anisotropic elliptic problems. In such a problem occurring in many models of physics, there exists a preferred direction corresponding to the direction of dominant diffusion along which the diffusion coefficient can be several orders larger than in the perpendicular one.

Thus, classical methods based on non-aligned stencils are particularly inefficient. Indeed, the latter are known to produce significant spurious numerical diffusion in the direction orthogonal to the dominant direction, and so to provide poor accuracy in describing possibly fast spatial variations in the directions across the main diffusion direction.

Motivated by the simulation of fusion plasmas in tokamak, the present scheme is designed in the finite-difference framework and standard Cartesian grids, but using interpolations aligned along the parallel direction. As recently proposed by Stegmeir and co-authors [21] to discretize the component of the differential operator in the direction parallel to the magnetic field line, the discretization is based on the Support Operator Method (SOM) that maintains the self-adjointness property of the parallel diffusion operator at the discrete level.

Under the single assumption of a divergence-free vector field that never vanishes to define the anisotropy directions, the present work introduces a detailed formulation of conservative discretizations of the parallel and perpendicular operators for non homogeneous systems. To make clearer and easier to write the implementation of such discretizations, the paper shows the construction of the stencils in 2D and for both curved and straight parallel field lines. The corresponding 2D numerical tests based on manufactured solutions show that all features of the aligned methods mentioned in the literature

Please cite this article in press as: J.A. Soler et al., A new conservative finite-difference scheme for anisotropic elliptic problems in bounded domain, J. Comput. Phys. (2019), https://doi.org/10.1016/j.jcp.2019.109093

are recovered, in particular the fact that the present method allows to drastically reduce the number of mesh points with respect to non-aligned approaches, this reduction becoming even more significant as the anisotropy is increased. Moreover, this new scheme brings new features with respect to the literature which are crucial to get more accurate and reliable solutions in most of realistic applications in fusion:

- The present scheme guarantees by construction the conservativity of the fluxes, not only along the parallel direction but also in the direction across the main diffusion direction. The conservativity at a discrete level has been also illustrated whatever the grid resolution in numerical tests involving an interpolation step.
- A method to deal with domain boundaries has been also proposed, which is compatible with the aligned discretization adopted for inner nodes. This method provides a much better accuracy than the classical approach based on far extrapolated ghost points along the dominant direction, with the paramount feature of allowing the same reduction in mesh points with aligned discretizations as in unbounded domains. In addition, this method allows to maintain the conservativity of the fluxes in bounded domains.

In conclusion, we are confident that this method brings new key and practical features to accurately simulate highly anisotropic problems in realistic configurations with boundaries, as long as this direction is known and stationary in time, or at least slowly varying with respect to major physical phenomena of interest in order to avoid frequent and costly remeshing. Numerical tests with non homogeneous anisotropy directions analytically defined show results that are rather robust, particularly at high parallel diffusion for which only aligned modes are not damped. However, the practical use of the method with very complicated magnetic topologies would remain challenging. Assuming that strong variations of the pitch angle are not too localized in the computational domain, the development of a multidomain approach able to allow the mesh distribution and the interpolation to change according to the mean value of the pitch angle in each subdomain could be a fruitful perspective of this work.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

The project leading to this publication has received funding from Excellence Initiative of Aix-Marseille University-A*MIDEX, a French “Investissements d’Avenir programme”. J.A. Soler thanks Aix-Marseille University-A*MIDEX and Centrale Marseille for his PhD grant. This work was supported by the EUROfusion - Theory and Advanced Simulation Coordination (E-TASC) and has received funding from the Euratom research and training programme 2019-2020 under grant agreement No 633053. The views and opinions expressed herein do not necessarily reflect those of the European Commission.

Appendix A. Study of the sensitivity of the elliptic problem to μ in a periodic domain

The following elliptic boundary value problem has been considered in this work with $\mu = 1$:

$$-\nabla \cdot \mathcal{K} \nabla T + \mu T = S \quad \text{in } \Omega = [0, 2\pi] \times [0, 2\pi],$$

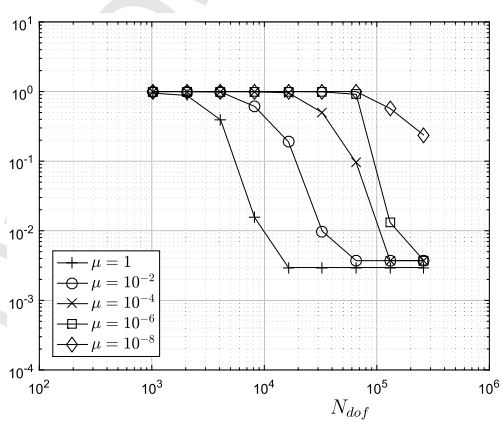


Fig. A.22. Plots of the $\mathcal{H}^1$-error with respect to the resolution N_{dof} for $\mu \in [10^{-6}, 1]$. Bi-periodic computational domain with $K_{b_\parallel} = 10^0$ and $K_{b_\perp} = 0$.

A non-zero value of μ allows us to consider a periodic computational domain, and so to separate the study of the discretization of the solution at the boundary with the one in the interior of the domain. Here, we show that the results presented in the paper are little sensitive to the value of μ. Fig. A.22 shows indeed that the $\mathcal{H}^1$-error converges whatever the value of μ, for $\mu \in [10^{-6}, 1]$. Obviously, when μ reaches near zero values, the problem above tends to become singular (Poisson's equation) in the periodic domain, and the resolution of the problem becomes much more demanding, which explains the increasing number of grid points needed to converge when μ decreases.

Appendix B. Representation of the solution with respect to the resolution

Results are discussed here with respect to the Nyquist-Shannon theorem [28,29], which provides the minimal resolution required to accurately represent the solution, i.e. $2m$ in each direction, where m is the highest wavenumber of the solution in this direction. The $\|T - T_a\|_{\mathcal{L}^2}$ and the $\|T - T_a\|_{\mathcal{H}^1}$ errors are plotted on Fig. B.23 for all numerical schemes.

For *non-aligned methods*, Fig. B.23a shows as expected that below the Nyquist-Shannon resolution (dotted line, $N_y = 2m_y$), the resolution is not fine enough to accurately discretize the solution and to decrease the errors. Spurious aliasing effects may eventually lead to a misleading small value of the $\mathcal{L}^2$ error, observed here for $N_{dof} = 2.88 \times 10^2$ with *Günter's method*. For larger resolutions ($N_y > 2m_y$) all the errors dominated by the discretization error in the y-direction decrease when increasing N_y. The minimal value is reached for a resolution corresponding to a perfect alignment of the grid with the solution (dashed line), i.e. for $N_y = N_x / \tan^{-1}\alpha$. In the present case, 3 grid points of the 9-point stencil used in *Günter's method* are exactly aligned with $\mathbf{b}$, and the stencil for the parallel Laplacian actually reduces to three points along the main diffusion direction. The parallel Laplacian of the aligned fluctuations is thus exactly zero at the discrete level, and aligned fluctuations are treated exactly. Beyond, the resolution in x being fixed, the discretization error in this direction becomes dominant and increases whatever the resolution used in the y-direction.

For *aligned methods* (Fig. B.23b), oriented stencils need an interpolation step in the y-direction to evaluate the parallel derivative introducing an additional discretization error related to the finite-difference scheme that can be large if the resolution is smaller than the Nyquist-Shannon resolution. For both resolutions in the x-direction, the error decreases when increasing the resolution in the y-direction. Oscillations of the error associated to local minima and maxima corresponding to resolutions for which the grids are aligned (minima) or the most misaligned (maxima) along the parallel diffusion direction. When using a finite-differences discretization, the interpolation error being proportional to $1/d_{b\|}^2 = 1/(\Delta x \cos^{-1}\alpha)^2$, where $d_{b\|}$ is defined in Eq. (25), it increases when N_x increases (i.e. $d_{\|}$ decreases) for the same number of grid points in the y-direction as shown on Fig. B.23b.

All these results justify the use of the $\mathcal{H}^1$-error in the analysis of the accuracy tests.

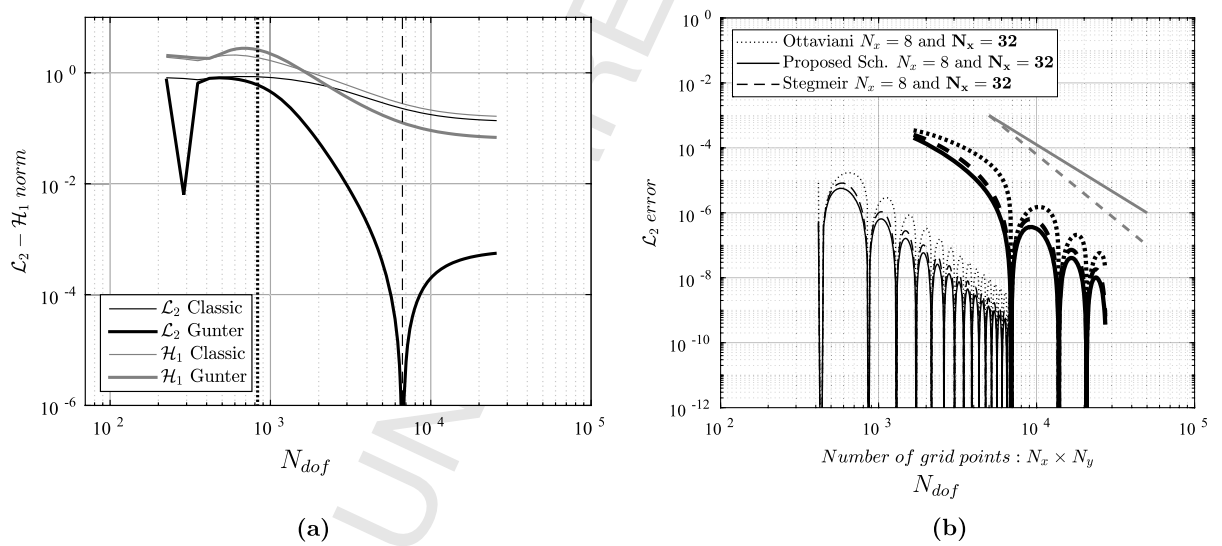


Fig. B.23. $\|T - T_a\|_{\mathcal{L}^2}$ and $\|T - T_a\|_{\mathcal{H}^1}$ errors when increasing resolution in the y-direction, $N_y \in [50, 250]$. (a) *Non-aligned methods*, $N_x = 32$. The dotted line corresponds to the minimal resolution prescribed by the Nyquist-Shannon theorem. The dashed line corresponds to the resolution for which the grid is perfectly aligned with the direction of the parallel diffusion. (b) *Aligned methods*, $N_x = 8$ (thin lines) and $N_x = 32$ (thick lines). T_a (Eq. (44)) is defined with $C_1 = C_3 = 0$, $C_2 = 1$ and with $m_y = 13$ and $m_{x,1} = 2$ leading to $\alpha = 8.75°$. With these values, the field remains constant along the parallel direction defined by $\mathbf{b} = (\cos\alpha, \sin\alpha, 0)$, while rapid variations can be observed in the perpendicular direction.

Please cite this article in press as: J.A. Soler et al., A new conservative finite-difference scheme for anisotropic elliptic problems in bounded domain, J. Comput. Phys. (2019), https://doi.org/10.1016/j.jcp.2019.109093

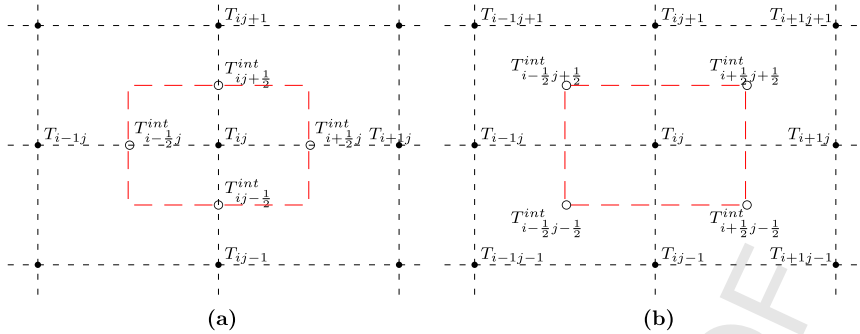


Fig. C.24. Examples of stencils. (a) The *classic method*. (b) *Günter's method*.

Appendix C. Discretization of the gradient in the $\mathcal{H}^1$-error

The $\mathcal{H}^1$-error is used in the paper to avoid misleading small values of the $\mathcal{L}^2$-error due to spurious aliasing effects introduced in the discretization of the anisotropic diffusion operator (see Appendix B).

The $\mathcal{H}^1$-error defines as:

$$\begin{aligned}\|T-T_a\|^2_{\mathcal{H}^1} &= \|T-T_a\|^2_{\mathcal{L}^2} + \|\nabla(T-T_a)\|^2_{\mathcal{L}^2} \\ &\geqslant \|T-T_a\|^2_{\mathcal{L}^2} + \|\nabla_x(T-T_a)\|^2_{\mathcal{L}^2} + \|\nabla_y(T-T_a)\|^2_{\mathcal{L}^2}\end{aligned}$$

requires the evaluation of the $\mathcal{L}^2$-error related to the gradients in each Cartesian direction x and y. Depending on the method used for the discretization, different stencils are used:

- For the *classical method*, gradients are evaluated by finite differences from grid points located in both x and y directions, Fig. C.24 (a). The gradients simply express as:

$$\nabla_x T_{i,j} \approx \frac{T_{i+1j} - T_{i-1j}}{2\Delta x}, \quad \nabla_y T_{i,j} \approx \frac{T_{ij+1} - T_{ij-1}}{2\Delta y}, \tag{C.1}$$

- For the *Günter's method*, the stencils involve the values of the function at the center of the surrounding cells Fig. C.24 (b) such as:

$$\begin{aligned}\nabla_x T_{i\,j} &\approx \frac{1}{2}\Big(\frac{T^{int}_{i+\frac{1}{2}\,j+\frac{1}{2}} + T^{int}_{i+\frac{1}{2}\,j-\frac{1}{2}}}{\Delta x} - \frac{T_{i-\frac{1}{2}\,j+\frac{1}{2}} + T^{int}_{i-\frac{1}{2}\,j-\frac{1}{2}}}{\Delta x}\Big) \\ \nabla_y T_{i\,j} &\approx \frac{1}{2}\Big(\frac{T^{int}_{i+\frac{1}{2}\,j+\frac{1}{2}} + T^{int}_{i-\frac{1}{2}\,j+\frac{1}{2}}}{\Delta y} - \frac{T^{int}_{i+\frac{1}{2}\,j-\frac{1}{2}} + T^{int}_{i-\frac{1}{2}\,j-\frac{1}{2}}}{\Delta y}\Big),\end{aligned} \tag{C.2}$$

  where T^{int} are evaluated from the nearest grid points as follows:

$$T^{int}_{i+\frac{1}{2}\,j+\frac{1}{2}} = \frac{T^{int}_{i+1\,j+1} + T^{int}_{i\,j+1} + T^{int}_{i+1\,j} + T^{int}_{i\,j}}{4} \tag{C.3}$$

- For the *aligned methods*, gradients in the x and y directions are obtained from the gradients evaluated in the parallel and perpendicular directions as detailed in the paper. Thus:

$$\nabla_\parallel T_{ij} \approx \frac{T^{int+} - T^{int-}}{(d_\parallel|^i_{i-1} + d_\parallel|^{i+1}_i)} \quad \nabla_\perp T_{ij} \approx \frac{\nabla_y T_{ij} - \nabla_\parallel T_{ij}\sin\alpha}{\cos\alpha}, \tag{C.4}$$

  and so (Fig. 9):

$$\nabla_x T_{i,j} \approx \nabla_\parallel T\cos(\alpha) - \nabla_\perp T\sin(\alpha), \quad \nabla_y T_{i,j} \approx \nabla_\parallel T\sin(\alpha) + \nabla_\perp T\cos(\alpha), \tag{C.5}$$

Appendix D. Resolution of the linear system

All the results presented in this paper require the inversion of a discrete matrix. The solution of the linear system Eq. (1) in a bi-periodic 2D domain is here calculated by the Matlab's *backslash* function for asymmetric sparse linear systems

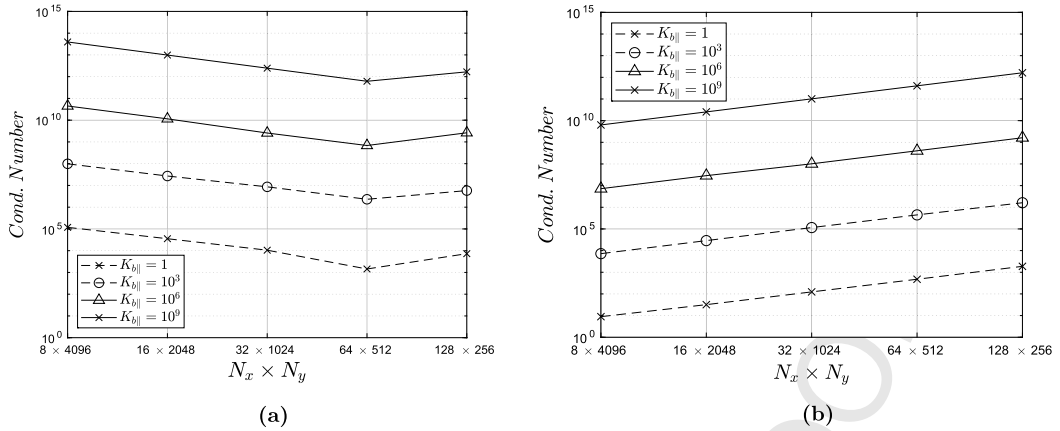


Fig. D.25. Plots of the values of the condition number for different grid points distribution and for different values of the parallel diffusion. N_x and N_y are varied keeping constant $N_{d.o.f.}$. The *Günter's method* (a) and the *present method* (b). $K_{b\perp} = 0$. The linear system Eq. (1) is considered here in a bi-periodic 2D domain.

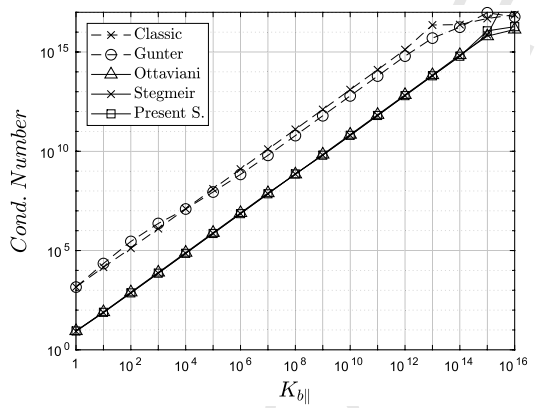


Fig. D.26. Plots of the condition number for all methods when varying $K_{b\|}$. The grids are 64×512 and 8×4096 for non-aligned and aligned methods, respectively. $K_{b\perp} = 0$. The linear system Eq. (1) is considered here in a bi-periodic 2D domain.

(UMFPACK [30]). The issue in the resolution of such system is mainly related to the high anisotropy possibly leading to an ill-conditioned discrete matrix, since the smallest eigenvalues, corresponding to eigenmodes in the null set of the parallel diffusion operator, are independent of $K_{b\|}$, whilst the largest eigenvalues scale as $K_{b\|}$ when $K_{b\|} \gg K_{b\perp}$ and $K_{b\|} \gg \mu$. We show here the evolution of the condition number calculated with Matlab for the different methods, and with respect to the grid distribution and the values of the parallel diffusion.

The Figs. D.25 plot the values of the condition number with respect to the distribution of the points between the x and y-directions, keeping constant $N_{d.o.f.}$ and for different values of $K_{b\|}$. The non aligned *Günter's method* and the aligned *present method* are considered. With the *present method*, the condition number is several orders below than with the *Günter's method* for a same parallel diffusion. The results show that the distribution of points also impacts the value of the condition number, but differently depending on the method, without it being possible to rigorously explain such a behavior.

Fig. D.26 plots now the condition number when varying the parallel diffusion, $K_{b\|}$. For all non-aligned methods, the 64×512 grid is considered, while a 8×4096 grid is chosen for all the aligned methods. Results show that for all methods the condition number grows linearly with $K_{b\|}$, with however a two orders lower value for the aligned methods. This is an important feature impacting the accuracy of the results when studying extreme values of $K_{b\|}$. The grid distribution has no impact here on the results for all the aligned methods.

References

[1] P. Tamain, H. Bufferand, G. Ciraolo, C. Colin, D. Galassi, Ph. Ghendrih, F. Schwander, E. Serre, The TOKAM3X code for edge turbulence fluid simulations of tokamak plasmas in versatile magnetic geometries, J. Comput. Phys. 321 (2016) 606–623.
[2] P. Tamain, Ph. Ghendrih, E. Tsitrone, Y. Sarazin, X. Garbet, V. Grand-Girard, J. Gunn, E. Serre, G. Ciraolo, G. Chiavassa, 3d modelling of edge parallel flow asymmetries, J. Nucl. Mater. 390–391 (2009) 347–350.
[3] M.V. Umansky, M.S. Day, T.D. Rognlien, On numerical solution of strongly anisotropic diffusion equation on misaligned grids, Numer. Heat Transf., Part B, Fundam. 47 (2005) 533–554.
[4] B. van Es, B. Koren, H.J. de Blank, Finite-difference schemes for anisotropic diffusion, J. Comput. Phys. 272 (2014) 526–549.

Please cite this article in press as: J.A. Soler et al., A new conservative finite-difference scheme for anisotropic elliptic problems in bounded domain, J. Comput. Phys. (2019), https://doi.org/10.1016/j.jcp.2019.109093

[5] P. Sharma, G.W. Hammett, Preserving monotonicity in anisotropic diffusion, J. Comput. Phys. 227 (2007) 123–142.
[6] I. Babuska, M. Suri, On locking and robustness in the finite element method, J. Comput. Phys. 29 (5) (1992) 1261–1293.
[7] S. Günter, Q. Yu, J. Krüger, K. Lackner, Modelling of heat transport in magnetised plasmas using non-aligned coordinates, J. Comput. Phys. 209 (2005) 354–370.
[8] L.G. Margolin, M. Shashkov, Finite volume methods and the equations of finite scale: a mimetic approach, Int. J. Numer. Methods Fluids 56 (2008) 991–1002.
[9] M. Shashkov, S. Steinberg, Support-operator finite-difference algorithms for general elliptic problems, J. Comput. Phys. 118 (1994) 131–151.
[10] J.M. Hyman, M. Shashkov, Approximation of boundary conditions for mimetic finite-difference methods, Comput. Math. Appl. 36 (5) (1998) 79–99.
[11] J.E. Morel, R.M. Roberts, M. Shashkov, A local support-operators diffusion discretization scheme for quadrilateral r-z meshes, J. Comput. Phys. 144 (1998) 17–51.
[12] K.V. Roberts, J.B. Taylor, Gravitational resistive instability of an incompressible plasma in a sheared magnetic field, Phys. Fluids 8 (2) (1965) 315–322.
[13] S. Cowley, R. Kulsrud, R. Sudan, Considerations of ion-temperature-gradient-driven turbulence, Phys. Fluids B 3 (10) (1991) 2767–2782.
[14] G.W. Hammett, M.A. Beer, W. Dorland, S.C. Cowley, S.A. Smith, Developments in the gyrofluid approach to tokamak turbulence simulations, Plasma Phys. Control. Fusion 35 (1993) 973–985.
[15] R.L. Dewar, A.H. Glasser, Ballooning mode spectrum in general toroidal systems, Phys. Fluids 26 (10) (1983) 3038–3052.
[16] B. Scott, Shifted metric procedure for flux tube treatments of toroidal geometry: avoiding grid deformation, Phys. Plasmas 8 (2) (2001) 447–458.
[17] S. Hamada, Hydromagnetic equilibria and their proper coordinates, Nucl. Fusion 2 (1962) 23–37.
[18] M. Ottaviani, An alternative approach to field-aligned coordinates for plasma turbulence simulations, Phys. Lett. 375 (2011) 1677–1685.
[19] F. Hariri, M. Ottaviani, A flux-coordinate independent field-aligned approach to plasma turbulence simulations, Comput. Phys. Commun. 184 (2013) 2419–2429.
[20] A. Arakawa, Computational design for long-term numerical integration of the equations of fluid motion: two-dimensional incompressible flow, part i, J. Comput. Phys. 135 (1997) 103–114.
[21] A. Stegmeir, D. Coster, O. Maj, K. Hallatschek, K. Lackner, The field line map approach for simulations of magnetically confined plasmas, Comput. Phys. Commun. 198 (2016) 139–153.
[22] A. Stegmeir, O. Maj, D. Coster, K. Lackner, M. Held, M. Wiesenberger, Advances in the flux-coordinate independent approach, Comput. Phys. Commun. 213 (2017) 111–121.
[23] Andrea Mentrelli, Claudia Negulescu, Asymptotic-preserving scheme for highly anisotropic non-linear diffusion equations, J. Comput. Phys. 231 (24) (2012) 8229–8245.
[24] Luis Chacon, Diego del Castillo-Negrete, C.D. Hauck, An asymptotic-preserving semi-lagrangian algorithm for the time-dependent anisotropic heat transport equation, J. Comput. Phys. 272 (2014) 719.
[25] Diego del Castillo-Negrete, Luis Chacon, Local and nonlocal parallel heat transport in general magnetic fields, Phys. Rev. Lett. 106 (2011) 195004, 05.
[26] K. Lipnikov, G. Manzini, M. Shashkov, Mimetic finite difference method, J. Comput. Phys. 257 (2014) 1163–1227.
[27] L.G. Margolin, M. Shashkov, Piotr K. Smolarkiewicz, A discrete operator calculus for finite difference approximations, Comput. Methods Appl. Mech. Eng. 187 (2000) 365–383.
[28] H. Nyquist, Certain topics in telegraph transmission theory, American Telephone and Telegraph Co. AIEE Winter Convention (1928).
[29] C.E. Shannon, A mathematical theory of communication, Bell Syst. Tech. J. 27 (1948) 379–423, 623–656.
[30] Timothy A. Davis, Algorithm 832: UMFPACK v4.3—an unsymmetric-pattern multifrontal method, ACM Trans. Math. Softw. 30 (2) (June 2004) 196–199.

Please cite this article in press as: J.A. Soler et al., A new conservative finite-difference scheme for anisotropic elliptic problems in bounded domain, J. Comput. Phys. (2019), https://doi.org/10.1016/j.jcp.2019.109093