Journal of Scientific Computing manuscript No.
(will be inserted by the editor)

The estimation of truncation error by τ-estimation for Chebyshev spectral collocation method

G. Rubio · F. Fraysse · J. de Vicente · E. Valero

Received: date / Accepted: date

Abstract In this paper we show how to accurately estimate the local truncation error of the Chebyshev spectral collocation method using τ-estimation. This method compares the residuals on a sequence of approximations with different polynomial orders. First, we focus the analysis on one-dimensional scalar linear and non-linear test cases to examine the accuracy of the estimation of the truncation error. Then, we show the validity of the analysis for the incompressible Navier-Stokes equations. First on the Kovasznay flow, where an analytical solution is known, and finally in the Lid Driven Cavity. We demonstrate that this approach yields a highly accurate estimation of the truncation error if the precision of the approximations increases with the polynomial order.

Keywords spectral methods · τ-estimation · truncation error · uncertainty estimator

Mathematics Subject Classification (2000) 65M70 · 65M15 · 65M50

1 INTRODUCTION

The estimation of numerical errors has been extensively used in numerical simulation [28]. Numerical errors provide valuable information about the quality of the solution [23,27]. Besides they are directly related to mesh adaptation [37,20]. One can find different approaches to the estimation of numerical errors depending on the numerical error of interest and the estimation method.

In recent years much work has been done on the estimation of the relative discretization error associated with functional outputs. This family of methods is called adjoint methodology and was first introduced by Venditti [34]. Adjoint

G. Rubio (Corresponding author), F. Fraysse, J. de Vicente and E. Valero
E.T.S.I.Aeronáuticos (Universidad Politécnica de Madrid)
Ciudad Universitaria, E-28040 Madrid, Spain
☎ +34-91-3366326
FAX +34-91-3366324
✉ g.rubio@upm.es

methodology permits accurate grid-induced corrections, specially for hyperbolic problems. However, its main drawback is its cost, as this approach requires the solution of the dual problem and usually the explicit storage of an embedded grid. In a high order context, this methodology has been recently used by Wang and Mavriplis [36] to estimate the error and perform mesh adaptation in a Discontinuous Galerkin (DG) method. Other authors have used this method to perform adaptation also in a DG framework. Examples of this can be found in the work of Leicht, Hartmann, Held and Prill [26] [25].

A different approach, very attractive due to its low computational cost, is the estimation of local errors. Moreover these methods provide an immediate strategy to perform h-refinement or p-refinement depending on the rate of convergence. Mavriplis [22,21] estimates the local discretization error by measuring the norm/energy associated to the different modes. This method has been also used in the last years by Casoni [10] in the scope of high order shock capturing schemes or by Rosenberg *et al.* [13] in the development of a object-oriented geophysical and astrophysical spectral-element adaptive refinement code. Also in the line of the estimation of local errors is the work by Wasberg and Gottlieb [11] where estimations of the local interpolation error are used to find optimal subdomain decompositions. In this case a wave-like behavior of the solution is supposed in the estimation, therefore it is only valid for problems whose solutions are mostly homogeneous over the whole computational domain.

Another alternative is to estimate the local truncation error by means of τ-estimation, first introduced by Brandt [7]. This technique has been used by Berger [3] in an adaptive Finite Differences (FD) method for the computation of a two-dimensional transonic NACA0012 Euler flow. Bernert [4] performed an extensive analysis of the accuracy of the method, extended later by Fulton [15]. In recent years Syrakos [33] successfully implemented τ-estimation for Finite Volumes (FV) discretization of the incompressible Navier-Stokes. More recently, Syrakos [32] also studied the efficiency of truncation error-based local refinement for the Navier-Stokes equations. Fraysse *et al.* [14] have extended those previous analysis to FV discretizations on any kind of meshes, with an interesting extension to τ-estimation using non-converged solutions. As can be seen, a lot of work has been done in the context of τ-estimation for low order methods. However, the authors are not aware of any devoted to its extension to high order methods.

This paper is dedicated to the analysis of the truncation error and the extension of the $\tau-$estimation methodology to spectral collocation methods.

As a preliminary step, a review of the behavior of the discretization and truncation errors in spectral collocation methods is made. Although the foundations of spectral theory can be found in text books such as: Canuto, Hussaini, Quarteroni and Zang [9], Boyd [6] or Kopriva [18] due to the minor diffusion of high order in contrast to low order methods, these expressions are not so commonly used and known, and they play a definitive role in our analysis.

The present paper is organized as follows. First, in Sec. 2, the mathematical formulation is derived, as well as the conditions to be fulfilled for an accurate estimation of the local truncation error. In Sec. 3, the previous analysis performed in Sec. 2 is validated and the truncation error for one-dimensional reference problems is estimated. Finally, in Sec. 4, more realistic configurations are addressed. Two reference problems in the Navier-Stokes equations are solved: the Kovasznay flow

[19], where an analytical solution is known, and the Lid Driven Cavity, by far the most used problem to test new algorithms in incompressible flows [16,8].

2 Mathematical Fundaments

In this section we formulate and solve the mathematical problem of the truncation error estimation for a collocation spectral method using τ-estimation. In the first subsection the mathematical problem is formulated. For the sake of self completeness of the paper we make a quick review of collocation spectral methods in subsections 2.2 and 2.3. In 2.4 using the theory presented, we develop a method for an accurate estimation of the truncation error. In Sec. 2.5 we show how to extend the method to several dimensions and systems of equations. Finally in 2.6 we analyze the computational cost of the method.

2.1 *Problem Formulation*

We start our analysis by considering a Partial Differential Equation (PDE)

$$\mathcal{L}u(x) = f(x) \tag{1}$$

where $\mathcal{L}$ is the partial differential operator, $u(x)$ is the exact solution of the problem and $f(x)$ is the forcing term of the equation. For simplicity we consider x extended over the domain $[-1, 1]$.

Let us consider the spectral collocation discretization of order N of Eq. 1

$$\mathcal{L}^N u^N = f^N \tag{2}$$

where $\mathcal{L}^N$ is the discretized partial differential operator using the collocation spectral method, u^N is the approximate solution and f^N the approximate forcing term. The PDE is discretized in a series of collocation points and solved there. The discretized PDE, Eq. 2, can be solved using an iterative method.

We define the current approximation of the solution (and not necessarily converged)

$$\tilde{u}^N = u^N - \epsilon_{it}^N \tag{3}$$

where ϵ_{it}^N is the iteration error.

We recall that the corresponding local truncation error is defined as follows:

$$\tau^N = \mathcal{L}^N u - f^N \tag{4}$$

the residual obtained by substituting the exact solution $u(x)$ onto the discretized PDE. Directly related to the local truncation error is the discretization error

$$\epsilon^N = u - u^N \tag{5}$$

which is the difference between the exact solution of the problem $u(x)$ and the approximate solution $u^N(x)$. The relationship between the local truncation error

and the discretization error is called the discrete discretization error transport equation (DETE, see Roy [28]), and reads

$$\tau^N = \mathcal{L}^N \epsilon^N \tag{6}$$

for linear operators and

$$\tau^N = \left.\frac{\partial \mathcal{L}^N}{\partial u^N}\right|_{u^N} \epsilon^N + \mathcal{O}((\epsilon^N)^2) \tag{7}$$

for non-linear operators.

In addition to the exact expression for the local truncation error, we introduce the next expression for the relative truncation error based on the low order relative truncation error used by Fraysse *et al.* in [14]:

$$\tau^N_{N+P} = \mathcal{L}^N \tilde{u}^{N+P} - f^N + \hat{I}^N_{N+P}(\mathcal{L}^{N+P}\tilde{u}^{N+P} - f^{N+P}) \tag{8}$$

where $\hat{I}^N_{N+P}$, the transfer operator of the iteration error, is defined as

$$\hat{I}^N_{N+P} = \mathcal{L}^N I_{N+P} (\mathcal{L}^{N+P})^{-1} \tag{9}$$

for linear operators and

$$\hat{I}^N_{N+P} = \left.\frac{\partial \mathcal{L}^N}{\partial u^N}\right|_{u^N} I^N_{N+P} \left(\left.\frac{\partial \mathcal{L}^{N+P}}{\partial u^{N+P}}\right|_{u^{N+P}} \right)^{-1} \tag{10}$$

for non linear operators.

Our goal is to use τ^N_{N+P} to estimate τ^N. For the case in which the solution is converged, $u^{N+P} = \tilde{u}^{N+P}$ and the second right hand side term of Eq. 8 can be neglected. Most of the analysis presented in the following sections is performed at convergence, and its extension to non-converged solutions is discussed in Sec. 3.1.3.

The following theorem provides the accuracy of the relative truncation error τ^N_{N+P} as estimator of the exact truncation error τ^N.

Theorem (Chebyshev Truncation Error Estimate) In the asymptotic range, for functions $u(x)$, $f(x)$ analytic in $[-1, 1]$ and with a regularity ellipse whose sum of semi-axes equals $e^\eta > 1$, the following expression holds

$$\tau^N_{N+P} = \tau^N + \mathcal{O}((N+P)^{1/2} N^{2l} e^{-(N+P)\eta}) \tag{11}$$

for linear opeartors and

$$\tau^N_{N+P} = \tau^N + \mathcal{O}(\max((N+P)^{1/2} N^{2l} e^{-(N+P)\eta}), \left(\epsilon^{N+P}_{it}\right)^2) \tag{12}$$

for non-linear operators. In both cases l is the highest order of derivation in the partial differential operator.

The proof of this theorem can be found in Sec. 2.4. It should be noticed that all the analysis have been made under the assumption of a value of P high enough

so the formal rate of convergence can be supposed. Thus, this analysis is only valid in the asymptotic range of convergence. It should be also noticed that in the non-linear expression, Eq. 12, there are contributions from the iteration error ϵ_{it} (derived from the use of a non-converged solution) that can be neglected for sufficiently converged solutions (so the error is dominated by the discretization error in the fine mesh, instead of the iteration error).

For the demonstration of the previously introduced theorem, some theory regarding spectral methods theory is required. For the sake of self completeness we present a quick review on spectral methods in Sec. 2.2. Using this theory we derive upper bounds for the truncation error in Sec. 2.3, which to the authors' knowledge have not been explicitly derived before. Further information about spectral methods can be found in [9], [6] or [18].

2.2 *Quick review on spectral collocation method*

We call interpolating polynomial of order N to the polynomial of the form:

$$I_N u(x) = \sum_{k=0}^{N} u(x_k) l_k(x) \tag{13}$$

where $u(x)$ is the function being approximated and $l_k(x)$ the k_{th} Lagrange polynomial. The precision of the method is set by the position of the nodes x_k. There is a direct relation among the set of nodes and the spectral basis used. In this work, we use the Chebyshev polynomials as the basis and the Chebyshev-Gauss-Lobatto (CGL) nodes. Nevertheless, the extension of the analysis to other combination of basis/nodes with spectral properties such as Fourier or Legendre is straightforward.

The interpolating polynomial is not exact for an arbitrary function u and a finite order N. We define the interpolating error ε_u^N as the difference between the function $u(x)$ and its interpolant of order N, $I_N u(x)$.

$$\varepsilon_u^N(x) = u(x) - I_N u(x) \tag{14}$$

It is important to make some remarks about Eq. 14. Firstly the definition is continuous. Secondly, it is observed that, according to the definition of Eq. 13, $\varepsilon_u^N(x_n) = 0$, which means that there is no error in the interpolation nodes. The interpolation error, ε_u^N, should not be confused with the previously defined discretization error ϵ^N. The latter is the difference between the exact solution and the numerical solution of Eq. 2, while the former is the difference between the exact solution and its spectral interpolant.

In a spectral collocation method the interpolating polynomial is used to approximate the differential operator. We present examples of this in Sec. 2.3. Under several conditions, the convergence of the method is exponential, also known in the literature as spectral convergence [9],[6],[18]. This means that, for a function $u(x)$, analytic in $[-1, 1]$ and with a regularity ellipse whose sum of semi-axes equals $e^\eta > 1$, in the asymptotic range of convergence (for N high enough) the following holds:

$$||u^{(l)} - (I_N u)^{(l)}||_{L_\infty(-1,1)} \leq C(\eta) N^{1/2} N^{2l} e^{-N\eta} \tag{15}$$

where l is the derivation order. For non analytic functions in $[-1, 1]$ the convergence deteriorates and it is known as algebraic. On the other hand, for entire functions the convergence of the method is even better and called super-geometric. In this case the error goes as $\mathcal{O}(N^{-N})$. Some authors [22] consider that both algebraic convergence and super-geometric convergence can be approximately predicted by exponential convergence with low or high values of η.

Eq. [15] can be used to derive an expression which defines the behavior of the local truncation error τ^N, in an equivalent way as the behavior of the truncation error in standard h schemes is $\mathcal{O}(h^p)$, where h is a typical mesh size and p the order of the scheme.

This upper bound depends both on the order of the interpolating polynomial N (or what is the same, the order of the spectral collocation method) and the order of the differential operator being discretized.

Finally, it has to be noticed that the discretization error ϵ^N of any problem solved using a spectral method is bounded as

$$\begin{aligned} ||u - u^N||_{L_\infty(-1,1)} &< C||u - I_N u||_{L_\infty(-1,1)} \\ ||\epsilon^N||_{L_\infty(-1,1)} &< C||\varepsilon_u^N||_{L_\infty(-1,1)} \end{aligned} \tag{16}$$

The above equation states that the discretization error is at most a constant away from the difference between the solution and the best polynomial approximation to the solution, which is a property of the spectral discretization [22]. We will use this expression here after to relate ϵ^N and ε_u^N.

Although the formulation used in the analysis of the error is continuous, in order to solve the numerical problem we need a discrete formulation of the problem. This discrete formulation (known as collocation) is obtained by evaluating the polynomial interpolants and its derivatives in the CGL nodes. For example, we can construct an interpolation matrix operator of order $N + P$ to transfer the values of an unknown function $f(x)$ from $N + P$ CGL nodes to N CGL nodes by:

$$F^N = I_{N+P}^N F^{N+P} = \sum_{k=0}^{N+P} f(x_k) l_k(x_j) \tag{17}$$

Where x_j are the CGL nodes of order N and x_k the CGL nodes of order $N + P$. The continuous expressions derived in the paper are always applicable to the discrete formulation, taking into account that the discrete formulation is no more than the continuous formulation evaluated in some points.

Another remark should be done about the notation followed in this work. The next two expressions are equivalent

$$\mathcal{L}^N u^{N+P} = \mathcal{L}^N I_{N+P}^N u^{N+P}. \tag{18}$$

In order to apply the discretized operator $\mathcal{L}^N$ to a solution of different order u^{N+P}, it is necessary to evaluate this solution in the CGL nodes of order N. In a discrete fashion this evaluation is seen as the interpolation represented by I_{N+P}^N. In this work we usually use the left hand side notation for compactness. It should also be noticed that the next convention for notation is followed:

$$f^N = I_N f \tag{19}$$

this means that f^N is the interpolating polynomial of order N of the function f.

In the next section we derive upper bounds for the local truncation error for two one-dimensional reference differential operators.

2.3 *Local truncation error analysis*

For illustration, in this section we derive the expressions which describe the behavior of the local truncation error in two reference problems. These problems are representative for the one-dimensional case and their results can be extrapolated to more complex scenarios.

– **Linear case** We consider the equation

$$u_{xx} = f \tag{20}$$

The problem is discretized by substituting u_{xx} by the second derivative of the interpolating polynomial of the unknown solution function $(I_N u^N)_{xx}$ and f by its interpolating polynomial $I_N f$. At this stage Dirichlet boundary conditions are considered. The boundary conditions can be imposed by substituting two equations for the exact value in the boundaries. More information about the discretization and how to impose boundary conditions can be found in [35]. Finally Eq. [20] becomes

$$(I_N u^N)_{xx} = I_N f \quad or \quad (u^N)_{xx} = f^N \tag{21}$$

Which in a discrete formulation is

$$D_N^2 U^N = F^N \tag{22}$$

where D^2 is the derived interpolator evaluated in the CGL nodes, U^N is the approximated solution in the CGL nodes and F^N is the forcing term in the CGL nodes. The truncation error of the differential operator is obtained by substituting the real solution in the discretized problem, Eq. [21],

$$\tau^N = (I_N u)_{xx} - I_N f \tag{23}$$

Contrary to the usual analysis in low order methods, we use a definition of τ^N in a continuos framework (with respect to the variable x). The discretization appears in the order of the polynomial N. By using the definition of the interpolating polynomial error Eq. [14] and the PDE definition Eq. [20], it can be deduced that

$$\begin{aligned} \tau^N &= u_{xx} - (\varepsilon_u^N)_{xx} - f + \varepsilon_f^N = \\ &= -(\varepsilon_u^N)_{xx} + \varepsilon_f^N \end{aligned} \tag{24}$$

And, by Eq. [15] and under the regularity assumptions defined above and for the asymptotic range of convergence

$$||(\varepsilon_u^N)_{xx}||_{L_\infty(-1,1)} \leq C(\eta) N^{1/2} N^{2\times 2} e^{-N\eta} \qquad (l=2) \tag{25}$$

and

$$||\varepsilon_f^N||_{L_\infty(-1,1)} \leq C(\eta) N^{1/2} e^{-N\eta}. \tag{26}$$

Finally, the value of η only depends on the position of the poles of f and u in the complex plane. Thus, assuming that the position of the poles of f and u are the same

$$||\tau^N||_{L_\infty(-1,1)} \leq C(\eta)N^{1/2}N^{2\times 2}e^{-N\eta}. \tag{27}$$

This formula describes the behavior of τ^N in the asymptotic range of convergence. As far as the discretization error is concerned, from Eq. 16 we know that its behavior is the same as for the spectral interpolating polynomial. So we can apply Eq. 15 under the assumptions already made there, to say that

$$||\epsilon^N||_{L_\infty} \leq CN^{1/2}e^{-N\eta}. \tag{28}$$

It is important to remark that this expression does not depend on the order of the problem but only on the regularity of the present functions $u(x)$ and $f(x)$. This is a remarkable difference with low order methods, where the truncation and discretization errors are of the same order of magnitude.

– **Non-linear case** The second problem considered is the steady state of the forced Burgers equation:

$$uu_x - u_{xx} = f. \tag{29}$$

There are several ways to discretize the non-linear term. Further information about the different methods which can be used can be found in [18]. We will use the straightforward discretization

$$I_N u \times (I_N u)_x - (I_N u)_{xx} = I_N f \tag{30}$$

Or in discrete form

$$U^N D_N U - D_N^2 U^N = F^N. \tag{31}$$

As before, we are interested in the calculation of the truncation error. If we use again the definition of the interpolating polynomial error Eq. 14 we have

$$\tau^N = (u - \varepsilon_u^N)(u_x - (\varepsilon_u^N)_x) - (u_{xx} - (\varepsilon_u^N)_{xx}) - (f - \varepsilon_f^N) \tag{32}$$

and neglecting the $\mathcal{O}({\varepsilon_u^N}^2)$ terms:

$$\tau^N = -u(\varepsilon_u^N)_x - u_x(\varepsilon_u^N) - (\varepsilon_u^N)_{xx} + \varepsilon_f^N \tag{33}$$

If we suppose that u and u_x are $\mathcal{O}(1)$, we can use Eq. 15 to estimate the order of magnitude of different terms involved, thus

$$\begin{aligned} ||\varepsilon_u^N||_{L_\infty(-1,1)} &\leq C(\eta)N^{1/2}e^{-N\eta} \\ ||(\varepsilon_u^N)_x||_{L_\infty(-1,1)} &\leq C(\eta)N^{1/2}N^{2\times 1}e^{-N\eta} \\ ||(\varepsilon_u^N)_{xx}||_{L_\infty(-1,1)} &\leq C(\eta)N^{1/2}N^{2\times 2}e^{-N\eta} \\ ||\varepsilon_f^N||_{L_\infty(-1,1)} &\leq C(\eta)N^{1/2}e^{-N\eta} \end{aligned} \tag{34}$$

As before, we have supposed u and f have the same poles. Therefore in the asymptotic range, the behavior of the truncation error τ^N is

$$||\tau^N||_{L_\infty(-1,1)} = CN^{1/2}N^{2\times 2}e^{-N\eta} \tag{35}$$

The same as before. As far as the discretization error is concerned, the result is the same that for the linear case as Eq. 16 does not depend on the order or the non-linearities of the problem solved, but only on the maximum derivative of the differential operator.

In this section we have derived upper bounds for the local truncation error of two one-dimensional operators, with three main conclusions. The first one is that the truncation error for spectral methods depends on the regularity of the functions involved in the problem (solution, forcing terms...). The second one states that the local truncation error is directly related to the maximum order of derivation of the differential operator. The third one is that the discretization error only depends on the regularity of the functions involved and the order of the method N.

With this, all the theory necessary for the proof of the Chebyshev truncation error estimate theorem has been presented. The demonstration is presented in the next section.

2.4 *Chebyshev truncation error estimate*

We are now in a position of proving the Chebyshev truncation error estimate theorem formulated in Sec. 2.1.

***Proof*:** We make a distinction between linear and non-linear differential operators, extending the analysis for non-converged solution.

- **Linear case**
  - **Converged solution**
    Firstly, we deduce the relation between the interpolant of order N for a given function $I_N u$ and the interpolant of order N of the interpolant of order $N+P$ of the same function $I_N(I_{N+P}u)$. This is:

$$I_N(I_{N+P}u) = I_N u - \sum_{k=0}^{N} \varepsilon_u^{N+P}(x_k) l_k(x) \tag{36}$$

which can be easily seen by calculating the difference between the two:

$$\begin{aligned} I_N(I_{N+P}u) - I_N u &= \sum_{k=0}^{N} (I_{N+P}u(x_k) - u(x_k)) l_k(x) \\ &= -\sum_{k=0}^{N} \varepsilon_u^{N+P}(x_k) l_k(x) = -I_N \varepsilon_u^{N+P} \end{aligned} \tag{37}$$

For the case in which the solution is converged, $u^{N+P} = \tilde{u}^{N+P}$ and the second right hand side term of Eq. 8 can be neglected. So, Eq. 8 reads:

$$\begin{aligned} \tau_{N+P}^{N} &= \mathcal{L}^N u^{N+P} - f^N = \\ &= \mathcal{L}^N (u - \epsilon^{N+P}) - f^N = \\ &= \mathcal{L}^N u - f^N - \mathcal{L}^N \varepsilon_u^{N+P} - \mathcal{L}^N \epsilon^{N+P} = \\ &= \tau^N - \mathcal{L}^N \varepsilon_u^{N+P} - \mathcal{L}^N \epsilon^{N+P} \end{aligned} \tag{38}$$

The term $\mathcal{L}^N \varepsilon_u^{N+P}$ appears because of the interpolation of the function u from the $N+P$ CGL nodes to the N CGL nodes. By Eq. 16 it can be

seen that the last two terms are of the same order of magnitude, and using Eq. 28 we can write

$$\frac{\epsilon^{N+P}}{\epsilon^N} = \mathcal{O}\left(\frac{(N+P)^{1/2}}{N^{1/2}} e^{-P\eta}\right) \tag{39}$$

so

$$\mathcal{L}^N \epsilon^{N+P} = \left(\mathcal{L}^N \epsilon^N\right) \mathcal{O}\left(\frac{(N+P)^{1/2}}{N^{1/2}} e^{-P\eta}\right) \tag{40}$$

and, as $\tau^N = \mathcal{L}^N \epsilon^N$ and $\tau^N = \mathcal{O}(N^{1/2} N^{2\ell} e^{-N\eta})$

$$\mathcal{L}^N \epsilon^{N+P} = \mathcal{O}\left((N+P)^{1/2} N^{2l} e^{-(N+P)\eta}\right) \tag{41}$$

finally

$$\tau^N_{N+P} = \tau^N + \mathcal{O}((N+P)^{1/2} N^{2l} e^{-(N+P)\eta}) \tag{42}$$

– **Non-converged solution**

Let us decompose the approximate solution $\tilde{u}^{N+P}$ such that $\tilde{u}^{N+P} = u - \epsilon^{N+P} - \epsilon_{it}^{N+P}$.

The Eq. 8 can be written

$$\begin{aligned}
\tau^N_{N+P} &= \mathcal{L}^N(u - \epsilon^{N+P} - \epsilon_{it}^{N+P}) - f^N + \hat{I}^N_{N+P}(\mathcal{L}^{N+P}\tilde{u}^{N+P} - f^{N+P}) = \\
&= \mathcal{L}^N u - f^N - \mathcal{L}^N \varepsilon_u^{N+P} - \mathcal{L}^N \epsilon^{N+P} - \\
&- \mathcal{L}^N \epsilon_{it}^{N+P} + \hat{I}^N_{N+P}(\mathcal{L}^{N+P}\tilde{u}^{N+P} - f^{N+P}) = \\
&= \tau^N - \mathcal{L}^N \varepsilon_u^{N+P} - \mathcal{L}^N \epsilon^{N+P} - \mathcal{L}^N \epsilon_{it}^{N+P} + \\
&+ \hat{I}^N_{N+P}(\mathcal{L}^{N+P}\tilde{u}^{N+P} - f^{N+P})
\end{aligned} \tag{43}$$

Two new terms due to the non-convergence of the solution have appeared: $-\mathcal{L}^N \epsilon_{it}^{N+P}$ and $\hat{I}^N_{N+P}(\mathcal{L}^{N+P}\tilde{u}^{N+P} - f^{N+P})$. If the solution is not converged, those terms can be of the same order of magnitude of the local truncation error. The transfer operator $\hat{I}^N_{N+P}$ is constructed in such a way that those terms cancel out. Indeed, from the definition of the iteration error (Eq. 3)

$$\mathcal{L}^{N+P}\tilde{u}^{N+P} - f^{N+P} = \mathcal{L}^{N+P}\epsilon_{it}^{N+P} \tag{44}$$

So in order to make those terms to cancel out

$$\hat{I}^N_{N+P}(\mathcal{L}^{N+P}\epsilon_{it}^{N+P}) = \mathcal{L}^N I^N_{N+P} \epsilon_{it}^{N+P} \tag{45}$$

the following relation must be fulfilled

$$\hat{I}^N_{N+P}\mathcal{L}^{N+P} = \mathcal{L}^N I_{N+P} \tag{46}$$

In this case, the expression for the truncation error estimation is

$$\tau^N_{N+P} = \tau^N - \mathcal{L}^N \varepsilon_u^{N+P} - \mathcal{L}^N \epsilon^{N+P} \tag{47}$$

which is the same obtained for the converged case, and so the error estimation for the converged case also holds here

$$\tau^N_{N+P} = \tau^N + \mathcal{O}((N+P)^{1/2} N^{2l} e^{-(N+P)\eta}) \tag{48}$$

– **Non-linear case**
  – **Converged solution**

As before,

$$\tau_{N+P}^{N} = \mathcal{L}^{N} u^{N+P} - f^{N}$$

Similar to the linear case, if we decompose $u^{N+P} = u - \epsilon^{N+P}$, we obtain

$$\tau_{N+P}^{N} = \mathcal{L}^{N}(u - \epsilon^{N+P}) - f^{N} \tag{49}$$

and linearizing

$$\begin{aligned}\tau_{N+P}^{N} &= \mathcal{L}^{N} u - f^{N} - \left.\frac{\partial \mathcal{L}^{N}}{\partial u^{N}}\right|_{u^{N+P}} \epsilon^{N+P} - \left.\frac{\partial \mathcal{L}^{N}}{\partial u^{N}}\right|_{u^{N+P}} \varepsilon_{u}^{N+P} + \mathcal{O}((\epsilon^{N+P})^{2}) \\ &= \tau^{N} - \left.\frac{\partial \mathcal{L}^{N}}{\partial u^{N}}\right|_{u^{N+P}} \epsilon^{N+P} - \left.\frac{\partial \mathcal{L}^{N}}{\partial u^{N}}\right|_{u^{N+P}} \varepsilon_{u}^{N+P} + \mathcal{O}((\epsilon^{N+P})^{2}).\end{aligned} \tag{50}$$

Finally, we have to evaluate the order of magnitude of the two last terms. If we use Eq. [28]

$$\frac{\epsilon^{N+P}}{\epsilon^{N}} = \mathcal{O}\left(\frac{(N+P)^{1/2}}{N^{1/2}} e^{-P\eta}\right)$$

it can be written

$$\left.\frac{\partial \mathcal{L}^{N}}{\partial u^{N}}\right|_{u^{N+P}} \epsilon^{N+P} = \left(\left.\frac{\partial \mathcal{L}^{N}}{\partial u^{N}}\right|_{u^{N+P}} \epsilon^{N}\right) \mathcal{O}\left(\frac{(N+P)^{1/2}}{N^{1/2}} e^{-P\eta}\right). \tag{51}$$

Now, using the relationship between the local truncation error and the discretization error $\tau^{N} = \left.\frac{\partial \mathcal{L}^{N}}{\partial u^{N}}\right|_{u^{N}} \epsilon^{N} + \mathcal{O}(\epsilon^{2})$ the order of magnitude is

$$\left.\frac{\partial \mathcal{L}^{N}}{\partial u^{N}}\right|_{u^{N+P}} \epsilon^{N+P} = \mathcal{O}\left((N+P)^{1/2} N^{2l} e^{-(N+P)\eta}\right) \tag{52}$$

from Eq. [16], the same is applicable to the other term. Therefore

$$\tau_{N+P}^{N} = \tau^{N} + \mathcal{O}\left((N+P)^{1/2} N^{2l} e^{-(N+P)\eta}\right). \tag{53}$$

  – **Non-converged solution**

For the non-converged case, by proceeding in an equivalent manner to the linear analysis, assuming that $\epsilon_{it}^{N+P} << u$

$$\begin{aligned}\tau_{N+P}^{N} = \tau^{N} &- \left.\frac{\partial \mathcal{L}^{N}}{\partial u^{N}}\right|_{u^{N+P}} \epsilon^{N+P} - \left.\frac{\partial \mathcal{L}^{N}}{\partial u^{N}}\right|_{u^{N+P}} \varepsilon^{N+P} - \\ &- \left.\frac{\partial \mathcal{L}^{N}}{\partial u^{N}}\right|_{u^{N+P}} \epsilon_{it}^{N+P} + \hat{I}_{N+P}^{N} \epsilon_{it}^{N+P} + \mathcal{O}(\max((\epsilon^{N+P})^{2}, (\epsilon_{it}^{N+P})^{2})).\end{aligned} \tag{54}$$

Now, in contrast to the linear case, deriving a transfer operator which eliminates the remaining terms does not necessarily ensure that the estimation will be accurate before any relaxation is performed. To perform the linearization present in the equation we have supposed $\epsilon_{it}^{N+P} << u$. Besides, the error performed depends on $(\epsilon_{it}^{N+P})^2$.
The correct transfer operator to eliminate the new source of error is

$$\hat{I}_{N+P}^{N} \left.\frac{\partial \mathcal{L}^{N+P}}{\partial u^{N+P}}\right|_{u^{N+P}} = \left.\frac{\partial \mathcal{L}^{N}}{\partial u^{N}}\right|_{u^{N+P}} I_{N+P}^{N} \tag{55}$$

And, the error for the truncation error estimate is

$$\tau_{N+P}^{N} = \tau^{N} + \mathcal{O} \max\left((N+P)^{1/2} N^{2l} e^{-(N+P)\eta}, \left(\epsilon_{it}^{N+P}\right)^2\right). \tag{56}$$

As in the local truncation error analysis, the result in the local truncation error estimate is the same for both the linear and the non-linear operators, but for the iteration error term. The precision of the estimation depends basically on the accuracy of the approximated solution $\tilde{u}^{N+P}$.

2.5 Extension to several dimensions and systems of equations

The analysis performed does not make any assumptions about the spatial dimensions of the problem, therefore the extension of the PDE to higher order space is straightforward. However, there are some distinguishing features when the collocation method is applied to higher dimensional spaces which need to be pointed out. For the sake of simplicity we limit our analysis to the two-dimensional problem.

The expansion to the two-dimensional problem decouples the problem in two spatial dimensions and, as a consequence, the convergence of the method is independent in both directions. So the convergence analysis can be made as a linear combination of two one-dimensional cases, as follows:

$$||\epsilon^{N_x N_y}||_{L_\infty} \leq C N_x^{1/2} e^{-N_x \eta_x} + C N_y^{1/2} e^{-N_y \eta_y} \tag{57}$$

for the discretization error and

$$||\tau^{N_x N_y}||_{L_\infty(-1,1)} \leq C N_x^{1/2} N_x^{2l} e^{-N_x \eta_x} + C N_y^{1/2} N_y^{2l} e^{-N_y \eta_y} \tag{58}$$

for the truncation error. This analysis allows different polynomial orders in each direction ($N_x \neq N_y$). Furthermore, the method here described can be applied to problems with different ratios of convergence in both spatial directions.

In a system of equations, as the number of equations and variables is increased, so does the number of errors which play part in the analysis. There is one discretization error per variable and one truncation error per equation. Additionally the convergence of each variable depends on its smoothness. For example in the equation

$$\begin{aligned} \frac{du}{dt} &= v + u_x \\ \frac{dv}{dt} &= u + v_x \end{aligned} \tag{59}$$

there are two discretization errors (ϵ_u and ϵ_v) each one with its rate of convergence (η_u and η_v) and two truncation errors obtained of substituting the exact values of u and v in both discretized equations. It can be easily seen that the behavior of the truncation error in this case is driven by a combination of both variables, as follows:

$$\begin{aligned} \tau_u^N &= \mathcal{O}(N^{1/2}e^{-N\eta_v}) + \mathcal{O}(N^{1/2}N^2e^{-N\eta_u}) \\ \tau_v^N &= \mathcal{O}(N^{1/2}e^{-N\eta_u}) + \mathcal{O}(N^{1/2}N^2e^{-N\eta_v}) \end{aligned} \tag{60}$$

All these aspects will be studied in the numerical experiments of Sec. 4 where the method will be applied to the two-dimensional incompressible Navier-Stokes equations.

2.6 *Actual costs of the estimation of truncation error by τ-estimation*

In this section the computational cost of the truncation error estimation by means of $\tau-$estimation is analyzed. The computational cost is highly dependent on the consideration of the method as an a-posteriori or an a-priori error estimator (using a converged or a non-converged solution for the estimation). If the method is used as an a-priori error estimator, the inversion of the Jacobian matrix in the fine mesh is required to treat the iteration error, which can be very expensive. So the computational cost of this inversion should be added to the whole computational cost of the algorithm. On the other hand, if the method is used as an a-posteriori error estimator, the converged solution in the fine mesh should be computed. The computational cost devoted to acquire a solution depends on the algorithm used to integrate the equations, so in this analysis, we consider that the problem has been already solved. Therefore only the cost of the estimation of the error *per se* is considered.

The memory requirements of the a-posteriori $\tau-$estimation method are negligible since this method only involves an interpolation of the solution to a coarse mesh, and the evaluation of this solution in the coarse operator. As far as the time requirements are concerned, it is necessary to analyze the algorithm. According to Eq. [8], the estimation of the truncation error by τ-estimation involves:

1. The computation of the solution of the problem in a fine grid
2. The interpolation of the solution to a coarse grid
3. The calculation of the residual when using the interpolated solution in the coarse discretized operator

The calculation of the residual (step three) involves several operations of complexity $\mathcal{O}(N)^2$ entailing a computational cost of the same order of magnitude of advancing one time step in an Euler explicit scheme. The interpolation of the solution (step two) involves an operation of complexity $\mathcal{O}(N \times (N+P))$.

Therefore, once the solution has been obtained in the fine grid $N+P$ (step one), the computational cost of estimating the truncation error for a coarser grid N is, at most, of the same order of magnitude of advancing one explicit time step. Moreover beyond this point, with a insignificant effort, an accurate estimation of the truncation error for all the coarser meshes $[1, N+P-1]$ can be obtained. In Section 4 the time cost of the method used in a real test case is shown.

3 DETAILED ANALYSIS ON REFERENCE PROBLEMS

Here, we validate the previous expressions derived for the approximation of the local truncation error, discretization error and the relative truncation error. The validation is performed in a framework of one-dimensional linear and non-linear problems, with particular focus on the non-converged solution.

3.1 *One-dimensional problems*

We consider the one-dimensional diffusion (linear) and diffusion-convection (non-linear, Burger's equation) equations with known exact solutions. Using those solutions the exact values of both the local truncation error and the discretization error can be computed.

The one-dimensional equations considered are

$$u_{xx} = f_1, \qquad uu_x - u_{xx} = f_2 \tag{61}$$

with the following test functions:

$$\begin{cases} f_1(x) = \dfrac{384x^9}{\left(2-x^4\right)^4} + \dfrac{288x^5}{\left(2-x^4\right)^3} + \dfrac{24x}{\left(2-x^4\right)^2} \\ u(-1) = u_{ex}(-1), \;\; u(1) = u_{ex}(1) \end{cases} \tag{62}$$

$$\begin{cases} f_2(x) = -\dfrac{384x^9}{\left(2-x^4\right)^4} - \dfrac{288x^5}{\left(2-x^4\right)^3} - \dfrac{24x}{\left(2-x^4\right)^2} + \dfrac{4x^3\left(\dfrac{32x^6}{\left(2-x^4\right)^3} + \dfrac{12x^2}{\left(2-x^4\right)^2}\right)}{\left(2-x^4\right)^2} \\ u(-1) = u_{ex}(-1), \;\; u(1) = u_{ex}(1) \end{cases} \tag{63}$$

which have the following exact solution:

$$u = \frac{4x^3}{(2-x^4)^2}. \tag{64}$$

This function fulfills the premises of the Chebyshev truncation error estimate formulated in Sec. 2.4. The function is analytic in the interval $[-1, 1]$, but has a pole outside this interval which makes the value of $\eta = -0.606$. The expression to calculate the value of η, being the position of the pole known can be found in Boyd [6]. This low value of η produces a slow convergence (although spectral) so we can analyze the different errors for a wide range of polynomial orders before reaching the machine error. Additionaly, the value of η (and so the position of the nearest pole to the interval $[-1, 1]$) is constant for u, f_1 and f_2 (which is the usual case). Both equations have been solved using the spectral collocation method described in the previous section. The steady state solution was reached using a simple iterative Euler relaxation scheme.

We focus the analysis on the validation of the theoretical expressions derived in the first section. We are specially interested in two aspects: the first one is the behavior of the local truncation and discretization error in spectral collocation

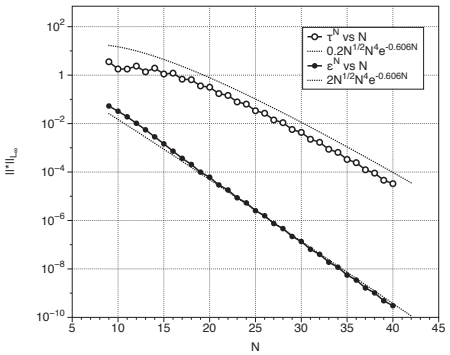

Fig. 1 Local truncation error and Discretization error for the linear case. Validation of Eq. 28 and Eq. 27

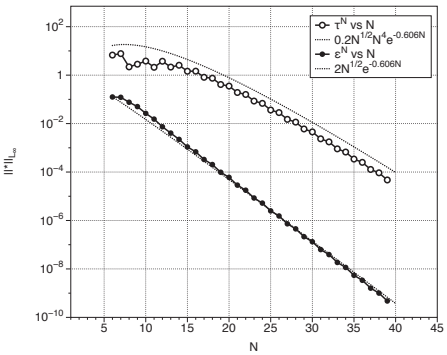

Fig. 2 Local truncation error and Discretization error for the non-linear case. Validation of Eq. 28 and Eq. 35

methods. The second one is the validation of the local truncation error estimation towards the relative truncation error. This estimation is analyzed for both converged and non-converged solutions.

3.1.1 Local truncation and discretization error behavior

Firstly, we validate the expressions for the discretization and the truncation errors (Eq. 28 and Eqs. 27, 35 respectively) derived in Sec. 2.3. Thus,

$$||\epsilon^N||_{L_\infty} \leq CN^{1/2}e^{-N\eta}$$
$$||\tau^N||_{L_\infty(-1,1)} \leq CN^{1/2}N^{2l}e^{-N\eta}$$

As already explained, both expressions only depend on the order of the differential operator (ℓ) and the regularity of the functions (η).

We have solved both problems for different orders of the method (different values of N) and calculated the exact values of local truncation and discretization errors. The results are shown in Figs. 1 and 2. The results obtained corroborate the theory. Two comments should be done. The first one is related to the oscillations seen in the curves. These are due to the modal nature of the method. When the order of the method is increased, we add new points but also new interpolating functions in order to solve the problem. When the introduced function is similar to the function we are approximating (for example a pair basis function for a pair approximated function) the decrease of the error is greater. Secondly, in the usual scenario the discretization error behaves better than the local truncation error. The N^{2l} term makes the discretization error to be lower than the local truncation error.

3.1.2 Local truncation error estimation - Converged solution

We check the validity of the Chebyshev truncation error estimate theorem derived in Sec. 2.4. Eq. 11 and Eq. 12 read

$$\tau^N_{N+P} = \tau^N + \mathcal{O}((N+P)^{1/2}N^{2l}e^{-(N+P)\eta})$$

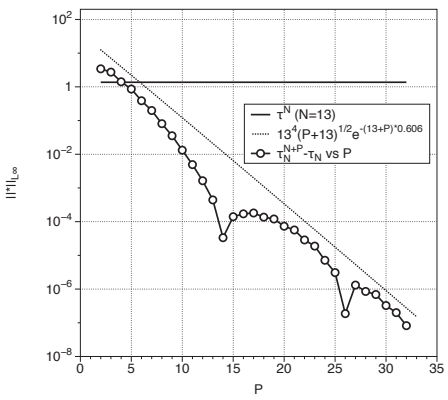

Fig. 3 Error in the truncation error estimation for the linear case. Converged solution N = 13. Validation of Eq. 11.

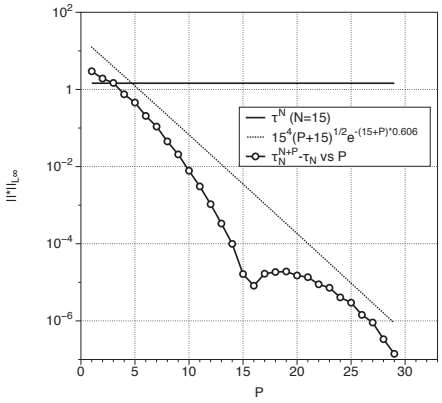

Fig. 4 Error in the truncation error estimation for the non-linear case. Converged solution N = 15. Validation of Eq. 12

$$\tau_{N+P}^N = \tau^N + \mathcal{O}(\max((N+P)^{1/2}N^{2l}e^{-(N+P)\eta}), \left(\epsilon_{it}^{N+P}\right)^2)$$

valid for both converged and non-converged linear and non-linear operators respectively. For the converged solution analysis presented here, $\left(\epsilon_{it}^{N+P}\right)^2 = 0$

In order to perform the validation, we have solved both the linear and the non-linear problems for different values of P with a fixed N. Then, we have used these solutions to calculate the relative truncation error τ_{N+P}^N using its definition of Eq. 8.

$$\tau_{N+P}^N = \mathcal{L}^N \tilde{u}^{N+P} - f^N + \hat{I}_{N+P}^N(\mathcal{L}^{N+P}\tilde{u}^{N+P} - f^{N+P})$$

which for the converged case simplifies to:

$$\tau_{N+P}^N = \mathcal{L}^N u^{N+P} - f^N$$

In order to do that, we have interpolated the solutions u^{N+P} from the finer grids $N+P$ to the coarser grid N, using the spectral interpolant of order $N+P$. In Figs. 3 and 4 we have represented the difference $||\tau^N - \tau_N^{N+P}||_{L_\infty}$ for a fixed value of N and increasing P. The real magnitude of the error made in the approximation has been compared to the one predicted by the theoretical analysis. The value of τ^N for fixed N has also been represented.

In the linear case, Fig. 3, we have fixed the value of $N = 13$. Then we have solved the problem using higher order $N + P$ with P ranging from $P = 1$ to $P = 32$. The behavior of the estimation error is the predicted but with periodical oscillations of amplitude N. These oscillations can be explained from the expression of the relative truncation error for converged solutions, Eq. 38,

$$\tau_{N+P}^N = \tau^N - \mathcal{L}^N \varepsilon_u^{N+P} + \mathcal{L}^N \epsilon^{N+P}$$

for the linear case and, Eq. 50

$$\tau_{N+P}^N = \tau^N - \left.\frac{\partial \mathcal{L}^N}{\partial u^N}\right|_{u^{N+P}} \epsilon^{N+P} - \left.\frac{\partial \mathcal{L}^N}{\partial u^N}\right|_{u^{N+P}} \varepsilon_u^{N+P} + \mathcal{O}((\epsilon^{N+P})^2)$$

for the non-linear case. In both equations there are terms in the error formula depending on the exactitude of the solution of order $N+P$: ϵ^{N+P} and on the accuracy of the interpolation of order $N+P$: ε^{N+P}. For $P = kN$ with $k = 1, 2, 3...$, the $N+1$ interpolation nodes of I_N coincide with the $N+P+1$ interpolation nodes of I_{N+P}. This entails injection instead of interpolation and a consequent diminution of the error. On the other hand, this error is also affected by modal effects which means that, apart from the low frequency oscillation seen at each N points, another high frequency oscillation is present depending on the quality of the function added with the increase of N.

In the non-linear case, Fig. 4, the polynomial order has been fixed to $N = 15$ while $P = [1, 29]$. The same behavior as for the linear case can be seen here.

3.1.3 Local truncation error estimation - Non-converged solution

In this section we present the result of using a non-converged solution in the local truncation error estimation.

In order to perform the validation, we have solved both the linear and the non-linear problems for fixed values of $N = 4$ and $P = 13$ and different levels of iteration error ϵ_{it}^N. Then we have used these solutions to calculate the relative truncation error τ_{N+P}^N using its definition Eq. [8].

$$\tau_{N+P}^N = \mathcal{L}^N \tilde{u}^{N+P} - f^N + \hat{I}_{N+P}^N (\mathcal{L}^{N+P} \tilde{u}^{N+P} - f^{N+P})$$

We present some results concerning the non-converged solution in Figs. 5 and 6. In both cases the truncation error estimation is presented with and without the correction term (second term of right hand side in Eq. [8]). This means, using the transfer operator to eliminate the error made by the iteration error. In the linear case, the truncation error estimation for the non-converged solution with correction is the same that the converged solution after one iteration. For the non-linear case, the decay rate of the error in the estimation is $\mathcal{O}((\epsilon_{it}^{N+P})^2)$, much better than without correction where the decay rate is $\mathcal{O}(\epsilon_{it}^{N+P})$. In both cases the result is the predicted by the theoretical analysis.

4 Numerical Experiments

The objective of this section is to validate the method in practical test cases. We complete our study by performing an analysis of the estimation of the truncation error on the two-dimensional incompressible Navier-Stokes equations.

Two different test cases are considered next: the Kovasznay flow and the Lid Driven Cavity (LDC). The first test case has been chosen because it has an analytical solution which incorporates non-linear effects. With this analytical solution in hand it is possible to prove the accuracy of the method in this complex problem. The second one is a test case whose main objective is to show how the method should be used in a real problem and so, no quantitative result is intended. Both test cases present smooth solutions, which is a requirement of the $\tau-$estimation method, in order to get accurate results.

The incompressible Navier-Stokes equations are solved using the artificial compressibility method [12], [31]. The main idea underlying this approach is to perturb

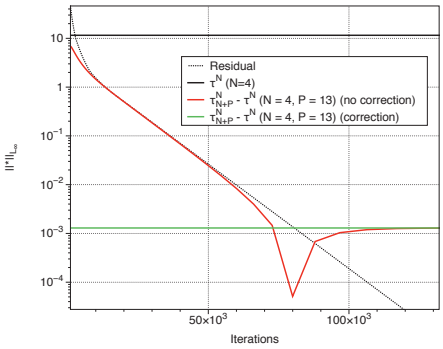

Fig. 5 Error in the truncation error estimation for the linear case. Behavior with the number of iterations for the non-converged case (with and without correction). N = 4, P = 13. Validation of Eq. [11]

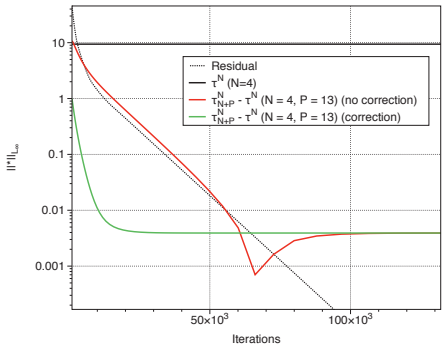

Fig. 6 Error in the truncation error estimation for the non-linear case. Behavior with the number of iterations for the non-converged case (with and without correction) N = 4, P = 13. Validation of Eq. [12]

the continuity equation by introducing a time derivative for the pressure in order to obtain a system of equations easier to solve, as follows:

$$\frac{\partial \mathbf{u}}{\partial t} + (\mathbf{u}\cdot\nabla)\mathbf{u} = -\nabla P + \frac{1}{Re}\nabla^2\mathbf{u} \tag{65}$$

$$\varepsilon\frac{\partial P}{\partial t} + \nabla\cdot\mathbf{u} = 0 \tag{66}$$

As before, Chebyshev spectral collocation technique has been chosen for the spatial discretization, while time advance has been performed using a semi-implicit Euler scheme. This semi-implicit numerical scheme treats convection terms explicitly but pressure terms implicitly. This approach avoids to solve a non-linear problem in each time step while preserves the stability properties of implicit solvers. The choice of an Euler method for time advancing is due to its simple implementation and its good stability properties when a steady solution is sought. This method has already been used in [35].

4.1 Kovasznay flow

The Kovasznay flow [19] is an analytical solution of the 2D steady-state incompressible Navier-Stokes equations that is similar to the laminar flow over a periodic array of cylinders. Since this flow incorporates non-linear effects (unlike Poiseuille flow), it is a good test for the full incompressible Navier-Stokes solution algorithm[17], [1], [2] and [29]. The analytical solution has the form:

$$\begin{aligned} u(x,y) &= 1 - e^{\lambda x}\cos(2\pi y) \\ v(x,y) &= \frac{\lambda}{2\pi}e^{\lambda x}\sin(2\pi y) \\ p(x,y) &= \frac{1}{2}(1 - e^{2\lambda x}) \end{aligned} \tag{67}$$

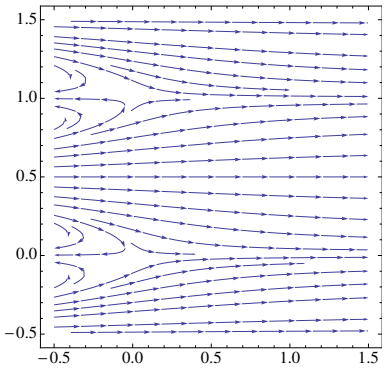

Fig. 7 Kovasznay Flow velocity field, Eq. 67, for Re = 40

where $\lambda = \frac{Re}{2} - \sqrt{\frac{Re^2}{4} + 4\pi^2}$. The velocity field for $Re = 40$ is shown in Fig. 7

The solution of the problem is analytic in the complex domain and so, the convergence of the method is supposed to be super-geometric (it converges as N^{-N}). However, as was stated in Sec. 2.2, even in this case it can be approximated by a geometric rate of convergence ($e^{-\eta N}$) with a high value of η.

Similar to the one-dimensional test cases of Sec. 3.1, we study the behavior of the local truncation and discretization errors. We pay special attention to the implications of the second dimension and the two extra equations/variables. Besides, we validate the local truncation error estimation towards the relative truncation error in a practical test case. It should be noticed that only converged solutions are considered in the following, and thus, the influence of the iteration error is not studied.

4.1.1 Local truncation and discretization error behavior

The accuracy of the method depends, as in the one-dimensional case, on the quality of the approximation of the solution. However in the two-dimensional case it is necessary to decouple the accuracy of the approximation in $x-$component and $y-$component, as it was explained in Sec. 2.5, as follows:

$$||\epsilon^{N_x N_y}||_{L_\infty} \leq C N_x^{1/2} e^{-N_x \eta_x} + C N_y^{1/2} e^{-N_y \eta_y}$$
$$||\tau^{N_x N_y}||_{L_\infty(-1,1)} \leq C N_x^{1/2} N_x^{2l} e^{-N_x \eta_x} + C N_y^{1/2} N_y^{2l} e^{-N_y \eta_y}$$

Besides, as a result of dealing with a system of equations, one error per equation (or per variable) should be taken into account. For example in this case, the truncation error in the first momentum equation reads:

$$\tau^N = \frac{1}{Re}\mathcal{O}(N^{1/2}N^4 e^{-N\eta_x^u}) + \mathcal{O}(N^{1/2}N^2 e^{-N\eta_x^p}) + \mathcal{O}(N^{1/2}N^2 e^{-N\eta_x^u}) + \mathcal{O}(N^{1/2}N^2 e^{-N\eta_y^u}). \quad (68)$$

The truncation error will be the sum of all the terms, and it will be driven by the biggest one, which means the one with the lowest value of η. Figs. 8 and 9 report the behavior of the local truncation error and the discretization error for different

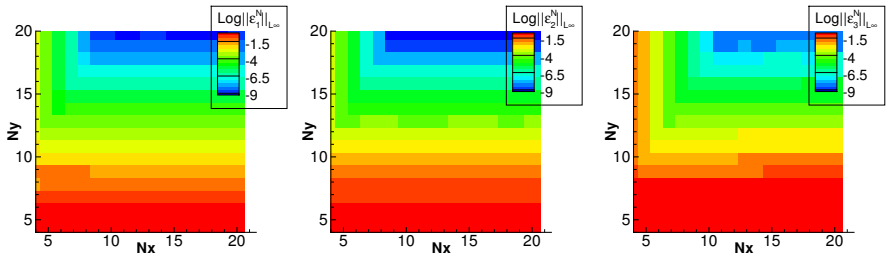

Fig. 8 Local discretization error for the Kovasznay flow. Validation of Eq. 57

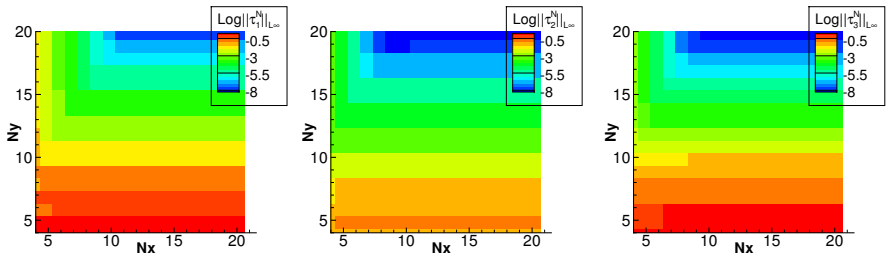

Fig. 9 Local truncation error for the Kovasznay flow. Validation of Eq. 58

values of N_x and N_y. In this case the component of the error which belongs to the y direction dominates for $N_x = N_y$. The reason of this is that the complexity of the solution is higher in y direction than in x direction. It can be also seen that the behavior in the three equations is similar.

In Sec. 2.5 it was stated that the convergence in each direction is independent in a multidimensional problem, furthermore the convergence in each direction is driven by the equations derived in the one dimensional analysis. In order to validate this, the discretization error for the horizontal velocity (u) and the truncation error in the first momentum equation, for fixed $N_x >> N_y$ and $N_y >> N_x$, are shown in Figs. 10 and 11 respectively. The behavior in these asymptotic cases is the same as in the one-dimensional case, as predicted. Besides it is important to remark that although the convergence of the method is super-geometric (due to the form of the solution), it is well approximated by the geometric convergence with high value of η.

4.1.2 Local truncation error estimation

The behavior of our estimator is also similar to the one-dimensional case. In this case, the problem has first been solved for $N_x = 6$ and $N_y = 6$. After that, we have solved it again for $N_x = 6 + P_x$ and $N_y = 6 + P_y$ for $P_x \in [1, 14]$ and $P_y \in [1, 14]$. Then, the higher order solution has been used to calculate the relative truncation error according to the definition. The error made in the approximation can be seen in Fig. 12.

As in the previous section, we have represented sections of Fig. 12 for different fixed values of P_x and P_y. The result, which can be seen in Fig. 13, proves the validity of the estimator for the 2D incompressible Navier-Stokes equations.

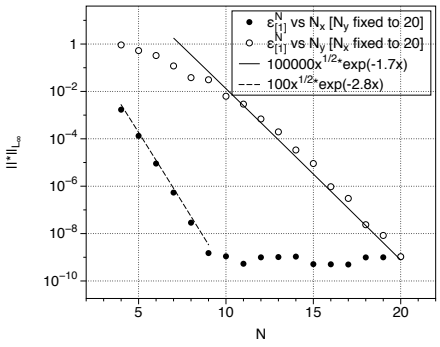

Fig. 10 Local Discretization error for the horizontal velocity u and fixed values of N_x and N_y. Validation of Eq. 57

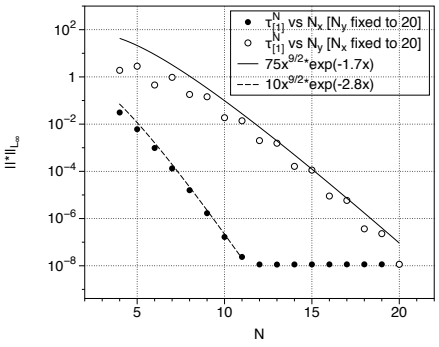

Fig. 11 Local exact truncation error for the first momentum equation and fixed values of N_x and N_y. Validation of Eq. 58

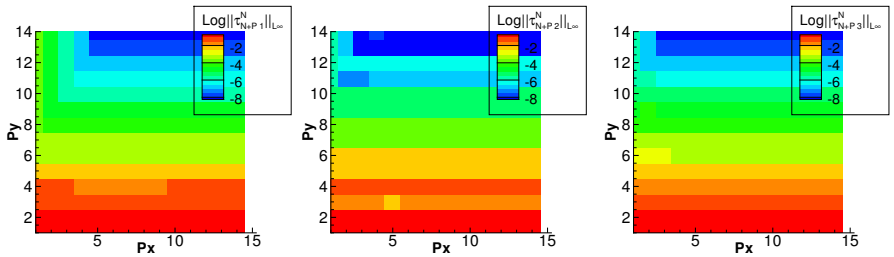

Fig. 12 Error in the truncation error estimation for the 2D linear case. Validation of Eq. 11 in 2D.

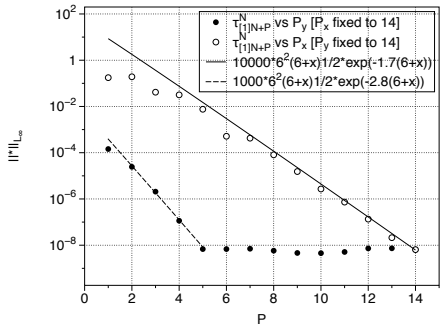

Fig. 13 Error in the truncation error estimation for the 2D linear case for fixed values of N_x and N_y. Validation of Eq. 11 in 2D.

4.2 Lid Driven Cavity

The suitability of the method for the 2D incompressible Navier-Stokes equations has been proved in Sec. 4.1 using the Kovasznay flow as test case. The objective of this section is to present a practical case in which the method can be used. We will pay special attention to describe the way the method should be used, the in-

formation it provides as well as its computational cost. As before, only converged solutions are considered in the following, and thus, the influence of the iteration error is not studied.

The Lid Driven Cavity (LDC) is the most used benchmark problem for testing new computational techniques for incompressible Navier Stokes solvers [16,8,5]. The test problem consists on a confined flow driven by one (usually the top) moving wall. In the original LDC problem the velocity of the top wall is constant. Unfortunately the abrupt change of the velocity from a non-zero constant value (top lid) to zero (side walls) results in a discontinuous flow field. This is a big drawback when used to test a spectral method because the order of convergence is deteriorated due to the discontinuity [5]. Particularly the $\tau-$estimation method assumes an smooth solution of the problem. One can find different approaches in the literature to adapt the LDC problem for spectral methods. Botella and Peyret [5] separates the most singular part obtaining the solution of the LDC. However, by far the most usual approach is to use the regularized driven cavity instead [24, 32,30,35], where the driving velocity is smoothed so that it vanishes (as well as its derivatives to fulfill the continuity equation) at the corners. In particular we use the same velocity distribution used in [32,17].

$$u(x) = -16x^2(1-x)^2 \tag{69}$$

which fulfills the already mentioned requirements.

In this test case we solve the LDC problem in a fine mesh and use the obtained solution to estimate the truncation error, towards $\tau-$estimation method, in all the coarser meshes. The estimations are used to analyze the accuracy of the obtained solution and the optimum procedure to get more accurate results. Finally some results are presented, using a more refined mesh, to validate the results of the fine mesh. In Table 1 a summary of the test case is shown.

Reynolds	Regularized velocity	Fine mesh	Coarse meshes	Fine mesh (validation)
1000	$-16x^2(1-x)^2$	20×20	$4/19 \times 4/19$	30×30

Table 1 Details on the test case of $\tau-$estimation in the LDC

4.2.1 Results

The results of the test case are presented here. In Figs. 14 and 15 the resulting solution in the 20×20 mesh is shown. In Fig. 16 the estimation of the truncation error in the coarser meshes $[4, 19] \times [4, 19]$ for each equation (first momentum, second momentum and continuity) is shown. In Fig. 17 the truncation error estimation for the continuity equation and fixed $N_y = 19$ and variable N_x and fixed $N_x = 19$ with variable N_y is shown. In Fig. 18 the truncation error estimation for the continuity equation obtained using the 20×20 and the 30×30 meshes is shown. As before, fixed values of N_y are used.

These results provide significant information about the problem solved. First important aspect, shown in Fig. 16, reveals that the complexity of the problem

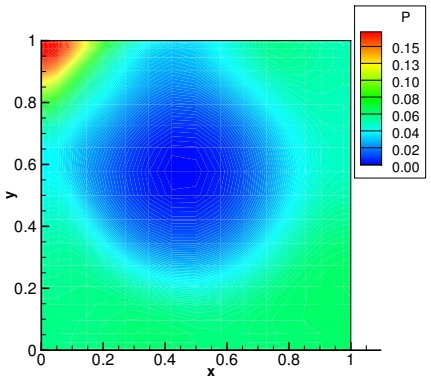

Fig. 14 Pressure contour in the 20×20 mesh

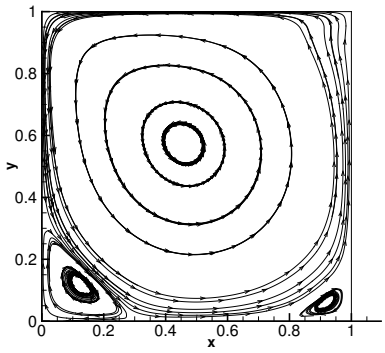

Fig. 15 Streamtraces in the 20×20 mesh

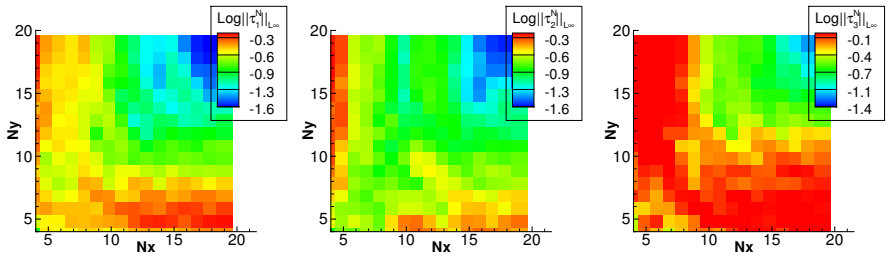

Fig. 16 Truncation Error estimation in the first moment equation, second momentum and continuity equations

is similar for the two spacial directions and the rate of convergence for all the equations is similar. This result contrasts with the one seen for the Kovasznay flow (Figs. 8, 9 and 12) where the complexity of the problem is higher in the y component. As the convergence of the method is independent in each spacial dimension, it is useful to analyze the results for N_x or N_y constant, in order to estimate these rates of convergence. An example of this estimation is shown in Fig. 17. Finally the accuracy of the solution in the 20×20 mesh can be approximated extrapolating the results in the coarser meshes. Beyond this point, if the obtained solution is not accurate enough, the polynomial order could be increased by using the estimation of the truncation error for each polynomial order.

Only with a validation purpose, different truncation error estimations using different fine meshes (20×20 and 30×30) are shown in Fig. 18. It should be remarked that this step is not required by the method.

4.2.2 Computational cost

In Sec. 2.6 the memory and time requirements of the method were analyzed. As it was exposed, the memory requirements of the method are negligible when used as an a-posteriori method. As far as the time requirements is concerned, the procedure is cheap, compared to the cost of finding an approximate solution of the Navier-Stokes equations.

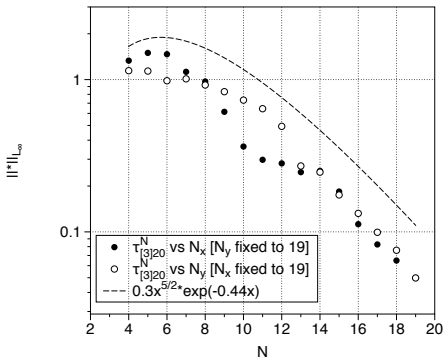

Fig. 17 Sectioned views of the Truncation Error estimation for the continuity equation

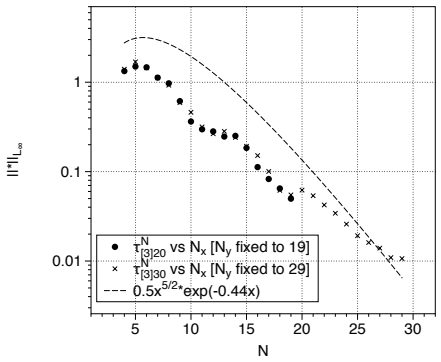

Fig. 18 Sectioned views of the Truncation Error estimation using differnent fine meshes for the continuity equation

In Table 2 we show a comparison between the time devoted to solve the problem and to perform the truncation error estimation. The latter is defined as the time consumed in the estimation of the truncation error in all the coarser meshes. Of course, these results depend on the numerical method used to integrate the equation and also on the Reynolds and polynomial order (through the CFL). In any case the time spent in the truncation error estimation is small compared to the one spent in solving the problem.

Polynomial Order $(N+P)$	Solve flow	$\tau^{N_x \times N_y}_{N_x+P_x \times N_y+P_y}$		No of estimations
10x10	55.6 s	0.9 s	1.59%	36
20x20	494.3 s	8.2 s	1.63%	256
30x30	2585.9 s	36.7 s	1.40%	676

Table 2 Time cost of solving the CFD problem compared to the truncation error estimation. These results where obtained in a MacBook Pro with a 2.4 GHz Intel Core 2 Duo processor and 4 GB of RAM memory.

5 CONCLUSIONS

Accurate estimations of the local truncation error have been successfully extended from low order methods to Chebyshev spectral collocation method. Conditions to ensure accurate estimations have been derived and verified numerically on the scalar Poisson equation, the Burgers equation and on the incompressible Navier-Stokes equations. In this approach, a converged solution is not assumed; thus, an analysis of the accuracy of the estimation has been performed within the iteration process to the steady state. The results demonstrated that for linear problems, it is possible to perform an accurate estimation at the first iteration and yields a robust a priori error estimator. For non-linear problems or if no special attention is provided, then the estimation is accurate as long as the magnitude of the iteration

error remains lower than the truncation error, although a method to mitigate the effect of the iteration error in the non-linear case have been derived. With an accurate estimation of the local truncation error in hand, several applications are natural, such as mesh generation and adaptation.

Acknowledgements

This research is supported by the European project ANADE (PINT-GA-2011-289428). Furthermore, the authors would like to thank Professor David A. Kopriva for his support, and for many invaluable discussions.

References

1. A. Ozcelikkale, C.S.: Least-squares spectral element solution of incompressible navier stokes equations with adaptive refinement. Journal of Computational Physics **231**, 3755–3769 (2012)
2. B. Cockburn, G.K., Tzau, D.: The local discontinuous galerkin method for the oseen equations. Mathematics of Computation **73**, 569593 (2003)
3. Berger, M.J.: Adaptive finite difference methods in fluid dynamics. Tech. rep., New York: Courant Institute of Mathematical Sciences, New York University (1987)
4. Bernert, K.: τ-extrapolation-theoretical foundation, numerical experiment, and application to navier-stokes equations. Siam Journal on Scientific Computing **18**, 460–478 (1997)
5. Botella, O., Peyret, R.: Benchmark spectral results on the lid-driven cavity flow. Computers & Fluids **27**(4), 421 – 433 (1998). DOI 10.1016/S0045-7930(98)00002-4. URL http://www.sciencedirect.com/science/article/pii/S0045793098000024
6. Boyd, J.: Chebyshev and Fourier spectral methods. Springer (1989)
7. Brandt, A., Livne, O.: Multigrid Techniques: 1984 Guide with Applications to Fluid Dynamics. SIAM (1984)
8. Burggraf, O.R.: Analytical and numerical studies of the structure of steady separated flows. Journal of Fluid Mechanics **24**, 113–151. DOI 10.1017/S0022112066000545. URL http://dx.doi.org/10.1017/S0022112066000545
9. C. Canuto M.Y. Hussaini, A.Q., Zang, T.: Spectral Methods in Fluid Dynamics. Springer-Verlag (1989)
10. Casoni, E.: Shock capturing for discontinuous galerkin methods. Ph.D. thesis, Universitat Politcnica de Catalunya (2011)
11. C.E. Wasberg, D.G.: Optimal decomposition of the domain in spectral methods for wave-like phenomena. SIAM J Sci Comput **22**(2), 617632 (2000)
12. Chorin, A.J.: A numerical method for solving incompressible flow problems. J. Comp. Phys. **2** (1967)
13. D. Rosenberg A. Fournier, P.F., Pouquet, A.: Geophysicalastrophysical spectral-element adaptive refinement (gaspar): Object-oriented h-adaptive fluid dynamics simulation. Journal of Computational Physics **215**(1), 59 – 80 (2006). DOI 10.1016/j.jcp.2005.10.031. URL http://www.sciencedirect.com/science/article/pii/S0021999105004791
14. Fraysse, F., De Vicente, J., Valero, E.: The estimation of truncation error by τ-estimation revisited. Journal of Computational Physics **231**, pp. 3457–3482 (2012)
15. Fulton, S.R.: On the accuracy of multigrid truncation error estimates. Electronic transactions on numerical analysis **15**, 29–37 (2003)
16. Ghia, U., Ghia, K., Shin, C.: High-re solutions for incompressible flow using the navier stokes equations and a multigrid method. Journal of Computational Physics **48**(3), 387 – 411 (1982). DOI 10.1016/0021-9991(82)90058-4. URL http://www.sciencedirect.com/science/article/pii/0021999182900584
17. Kondaxakis, D., Tsangaris, S.: A weak legendre collocation spectral method for the solution of the incompressible navier stokes equations in unstructured quadrilateral subdomains. Journal of Computational Physics **192**(1), 124 – 156 (2003). DOI 10.1016/S0021-9991(03)00350-4. URL http://www.sciencedirect.com/science/article/pii/S0021999103003504

18. Kopriva, D.A.: Implementing Spectral Methods for Partial Differential Equations: Algorithms for Scientists and Engineers. Scientific Computation. Springer Science+Business Media B.V. (2009)
19. Kovasznay, L.I.G.: Laminar flow behind a two-dimensional grid. Proc. Camb. Philos. Soc. **44**, 5862 (1948)
20. Löhner, R.: Mesh adaptation in fluid mechanics. Engineering Fracture Mechanics **50**(56), 819 – 847 (1995). DOI 10.1016/0013-7944(94)E0062-L. URL http://www.sciencedirect.com/science/article/pii/0013794494E0062L
21. Mavriplis, C.: Nonconforming discretizations and a posteriori error estimators for adaptive spectral element techniques. Ph.D. thesis, Massachusetts Institute of Technology (1989)
22. Mavriplis, C.A.: Adaptive mesh strategies for the spectral element method. Computer Methods in Applied Mechanics and Engineering **116**(14), 77 – 86 (1994). DOI 10.1016/S0045-7825(94)80010-3. URL http://www.sciencedirect.com/science/article/pii/S0045782594800103
23. Oberkampf, W.L., Roy, C.J.: Verification and Validation in Scientific Computing. Cambridge University Press (2010)
24. Peyret, R., Taylor, T.: Computational methods for fluid flow. Springer series in computational physics. Springer-Verlag (1983). URL http://books.google.com.ar/books?id=hZZRAAAAMAAJ
25. R. Hartmann, J.H., Leicht, T.: Adjoint-based error estimation and adaptive mesh refinement for the rans and k turbulence model equations. Journal of Computational Physics **230**(11), 4268 – 4284 (2011). DOI 10.1016/j.jcp.2010.10.026. URL http://www.sciencedirect.com/science/article/pii/S0021999110005826. Special issue High Order Methods for CFD Problems
26. R. Hartmann J. Held, T.L., Prill, F.: Discontinuous galerkin methods for computational aerodynamics 3d adaptive flow simulation with the dlr padge code. Aerospace Science and Technology **14**(7), 512 – 519 (2010). DOI 10.1016/j.ast.2010.04.002. URL http://www.sciencedirect.com/science/article/pii/S1270963810000441
27. Roache, P.J.: Verification and validation in computational science and engineering. Hermosa Publishers (1998)
28. Roy, C.J.: Review of discretization error estimators in scientific computing. In: AIAA Paper (2010)
29. Shah, A., Yuan, L., Khan, A.: Upwind compact finite difference scheme for time-accurate solution of the incompressible navier stokes equations. Applied Mathematics and Computation **215**(9), 3201 – 3213 (2010). DOI 10.1016/j.amc.2009.10.001. URL http://www.sciencedirect.com/science/article/pii/S0096300309008947
30. Shen, J.: Dynamics of regularized cavity flow at high reynolds numbers. Applied Mathematics Letters **2**(4), 381 – 384 (1989). DOI 10.1016/0893-9659(89)90093-1. URL http://www.sciencedirect.com/science/article/pii/0893965989900931
31. Shen, J.: Pseudo-compressibility methods for the unsteady incompressible navier stokes equations. Beijing Symposium on Nonlinear Evolution Equations and Infinite Dynamical Systems (1997)
32. Syrakos, A., Efthimiou, G., Bartzis, J.G., Goulas, A.: Numerical experiments on the efficiency of local grid refinement based on truncation error estimates. Journal of Computational Physics **231**(20), 6725 – 6753 (2012). DOI 10.1016/j.jcp.2012.06.023. URL http://www.sciencedirect.com/science/article/pii/S0021999112003385
33. Syrakos, A., Goulas, A.: Finite volume adaptive solutions using simple as smoother. International journal for numerical methods in fluids **52**, 1215–1245 (2006)
34. Venditti, D.A., Darmofal, D.L.: Adjoint error estimation and grid adaptation for functional outputs: Application to quasi-one- dimensional flow. Journal of Computational Physics **164**, 204–227 (2000)
35. de Vicente, J.: Spectral multi-domain methods for the global instability analysis of complex cavity flows. Ph.D. thesis, Polytechnic University of Madrid (2010)
36. Wang, L., Mavriplis, D.J.: Adjoint-based hp adaptive discontinuous galerkin methods for the 2d compressible euler equations. Journal of Computational Physics **228**(20), 7643 – 7661 (2009). DOI 10.1016/j.jcp.2009.07.012. URL http://www.sciencedirect.com/science/article/pii/S0021999109003854
37. Zienkiewicz, O.C.: The background of error estimation and adaptivity in finite element computations. Computer Methods in Applied Mechanics and Engineering **195**(46), 207 – 213 (2006). DOI 10.1016/j.cma.2004.07.053. URL http://www.sciencedirect.com/science/article/pii/S0045782505000824. Adaptive Modeling and Simulation