Contribution à l'analyse numérique de quelques problèmes en chimie quantique et mécanique.

Rachida Chakir

Université Pierre et Marie Curie – Paris 6 Laboratoire Jacques-Louis Lions – UMR 7598

Contribution à l'analyse numérique de quelques problèmes en chimie quantique et mécanique

THÈSE DE DOCTORAT

présentée et soutenue publiquement le 30 novembre 2009 par

Rachida Chakir

pour l'obtention du

Doctorat de l'université Pierre et Marie Curie – Paris 6

Spécialité Mathématiques Appliquées

Devant le jury composé de :

Éric CANCÈS	Examinateur
Bruno DESPRÉS	Examinateur
François JOLLET	Examinateur
Yvon MADAY	Directeur de thèse
Endre SÜLI	Rapporteur

Après avis des rapporteurs : Jean-Luc GUERMOND
Endre SÜLI

École Doctorale de Sciences Mathématiques de Paris Centre UFR 929 - Mathématiques

Remerciements

J'aimerais en premier lieu exprimer ma reconnaissance envers mon directeur de thèse, Yvon Maday, pour la confiance qu'il m'a accordée, en me permettant d'obtenir un financement sans lequel cette thèse n'aurait pas eu lieu. J'ai énormément appris durant ces trois années, j'ai pu découvrir un nouveau monde, celui de la recherche. L'aide qu'il m'a apportée ainsi que sa patience m'ont été très précieuses dans l'accomplissement de ce travail.

Je souhaiterais remercier mes rapporteurs Jean-Luc Guermond et Endre Süli pour le temps qu'ils ont accordés à la lecture de cette thèse et à l'élaboration de leur rapport. L'intérêt qu'ils ont porté à mes travaux ainsi que leurs critiques ont permis d'améliorer ce mémoire.

Un grand merci également à Eric Cancès pour ses conseils avisés durant cette thèse et pour sa présence dans mon jury. C'est également avec plaisir que je remercie François Jollet et Bruno Després d'avoir accepté de faire parti de mon jury.

Je tiens à témoigner ma reconnaissance à Edwige Godlewski pour son soutien et sa disponibilité dans les moments de doute.

Pour leur gentillesse et leur disponibilité, je remercie Danièle Boulic, Liliane Ruprecht, Florence Saidani, Salima, et Christian David. C'est également avec plaisir que je remercie Khashayar, Antoine Le Hyaric et Fréderic Hecht pour leur aide lors de mes déboires informatiques.

Je n'oublie évidemment pas mes amis et camarades du LJLL avec lesquels j'ai partagé ces trois dernières années.

Je remercie tout particulièrement Alexandra C., qui a su être présente à tout instant. Son soutien, toutes ces heures passées à me relire et corriger mes "perles" orthographiques et grammaticales et ses remarques ont été autant de mains tendues. Je pense également à Nicolas L., l'expert en Gnuplot et correcteur suppléant. Je ne pourrais jamais assez vous remercier pour votre aide (j'espère que vous avez gardé vos stylo rouge, car cette page doit comporter de beaux spécimens).

Un grand merci également à Laurent, Nicole et Ulrich pour leur nombreux conseils avisés.

Merci aux anciens et nouveaux du bureau 3D18, Joelle (tu nous manques, reviens vite), Paulo et Raphael. Mais aussi à Maya (promis je ferais moins de bruit et mon bordel restera de mon coté du bureau), Sépideh, Matthieu L., Giacomo, Tina, Alexandra F. et Joanna. Quant à nos voisins de paliers, je remercie Evelyne (pour ses réponses à mes questions mêmes les plus stupides), Mathieu G. (ils sont trop bons tes gâteaux), Alexis (les tiens aussi, la relève est assurée), Benjamin A., Pierre, Jean-Marie et Mouna.

Je n'oublie pas Julie, Cuc, Benjamin B., Vincent, Thomas, Etienne, Rym, Noura, les autres Nicolas, Bawer, Filipa, Céline et J.F.

Merci à vous tous, chacun à votre façon vous avez contribué à l'accomplissement de cette thèse.

Je termine ces remerciements, en pensant à mes parents qui m'ont toujours soutenue et encouragée.

Table des matières

Introduction et présentation des résultats

L'objet principal de cette thèse est une contribution à l'analyse numérique de problèmes de valeurs propres non linéaires, comme on peut en trouver en chimie quantique. La résolution de ces problèmes étant très coûteuse, l'idée est de proposer de nouvelles méthodes permettant de simplifier la résolution de ce type de problèmes et ainsi diminuer le coût total de calcul. L'analyse numérique est nécessaire pour comprendre si l'impact positif sur le coût de total n'a pas de mauvaise conséquence sur la précision des résultats.
On s'est aperçu que l'analyse numérique de discrétisation classique n'était pas entièrement faite, et surtout elle n'était pas optimale. Il a fallu préalablement compléter les travaux existant sur les estimations d'erreur *a priori*, afin d'obtenir des résultats équivalents à ceux connus dans le cas de problèmes aux valeurs propres linéaires. Les deux premiers chapitres sont consacrés à l'analyse numérique de ces problèmes aux valeurs propres non linéaires, ainsi que l'effet de l'intégration numérique.
Dans le chapitre 3, ces résultats ont été utilisés pour la mise en œuvre et l'analyse numérique de nouveaux *schémas à deux grilles* pour l'approximation de problèmes aux valeurs propres non linéaires.
Dans une dernière partie, on propose d'adapter ce type de méthode de *sous-grilles*, pour une utilisation originale, associée à la méthode des bases réduites pour la résolution de problèmes elliptiques avec second membre.

Quelques résultats d'estimations *a priori* dans le cas de problèmes aux valeurs propres linéaires

De nombreuses applications physiques et mécaniques, nécessitent l'approximation des valeurs propres et des vecteurs propres de problèmes elliptiques, ayant des conditions aux limites. L'analyse numérique de ces problèmes linéaires a été amplement étudiée et développée, et plus particulièrement dans le cas de la méthode des éléments finis [2, 45, 50]. La présentation de ces principaux résultats est faite dans le même esprit que celle qui sera proposée dans la suite de la thèse, pour le cas de problèmes non linéaires.

Soit Ω un ouvert borné de $\mathbb{R}^d$ ($d = 2$ ou 3) à frontière « régulière ». On s'intéresse au problème suivant :

Trouver $u \in H_0^1(\Omega)$ et $\lambda \in \mathbb{R}$ tels que

$$\begin{cases} a(u,v) = \lambda \displaystyle\int_\Omega uv, & \forall v \in H_0^1(\Omega) \\ \displaystyle\int_\Omega u^2 = 1, \end{cases} \tag{1}$$

où a est une forme bilinéaire, symétrique, continue et coercive sur $H_0^1(\Omega)$.
Les valeurs propres de ce problème forment une suite croissante tendant vers $+\infty$

$$0 < \lambda_1 \leq \lambda_2 \leq \cdots \leq \lambda_m \leq \cdots ,$$

et les fonctions propres associées u_m sont bornées dans $H^2(\Omega)$ (en supposant que la forme bilinéaire a et la frontière du domaine Ω sont suffisamment régulières, par exemple si Ω est convexe).
Il s'agit maintenant d'approcher numériquement les couples d'élements propres (λ, u). On introduit un sous-espace de $H_0^1(\Omega)$ de dimension finie, de type éléments finis $(K,\ P_K,\ \sum_K)$, noté X_h^k tel que :

$$X_h^k = \{v \in H_0^1(\Omega),\ \forall\, K \in \mathcal{T}_h,\ v_{|K} \in \mathbb{P}_k(K)\}.$$

$\mathcal{T}_h$ représente une famille régulière de triangulation de Ω, le paramètre de discrétisation h est défini par $h = \max\limits_{T\in\mathcal{T}_h} h_T$ où h_T est le diamètre de T (c'est à dire la longueur du plus grand coté). On rappelle qu'une famille de triangulation est dite régulière si elle vérifie les hypothèses suivantes :

- pour tout h, $\overline{\Omega}$ est égale à l'union de tous les éléments de $\mathcal{T}_h$,
- l'intersection de deux éléments distincts est vide, ou un sommet, ou une arête entière ou une face entière,
- il existe une constante σ indépendante de h telle que $\forall T \in \mathcal{T}_h$, $\sigma_T \leq \sigma$ avec $\sigma_T = \frac{h_T}{\rho_T}$, et ρ_T le diamètre de la boule inscrite dans T.

L'approximation de Galerkin de (1) sur X_h^k, s'écrit alors :

Trouver $u_h \in X_h^k$ et $\lambda_h \in \mathbb{R}$ tels que

$$\begin{cases} a(u_h, v_h) = \lambda_h \displaystyle\int_\Omega u_h v_h, \qquad \forall v_h \in X_h^k \\ \displaystyle\int_\Omega u_h^2 = 1. \end{cases} \tag{2}$$

Ce problème admet $N_h (= \dim X_h^k)$ valeurs propres positives qui forment une suite croissante :

$$0 < \lambda_{1,h} \leq \lambda_{2,h} \leq \cdots \leq \lambda_{m,h} \leq \cdots \leq \lambda_{N_h,h}.$$

Soit $\mathcal{V}_m$ l'ensemble des sous-espaces de $H_0^1(\Omega)$ et $\mathcal{V}_{m,h}$ l'ensemble des sous-espaces de X_h^k, tous deux de dimension m. Alors, d'après le Principe du Min-Max, la $m^{\text{ème}}$ valeur propre λ_m du problème (1) et la $m^{\text{ème}}$ valeur propre $\lambda_{m,h}$ du problème (2) sont données par

$$\begin{aligned} \lambda_m &= \min_{E_m\in\mathcal{V}_m} \max_{\substack{v\in E_m \\ v\neq 0}} \frac{a(v,v)}{\|v\|_{L^2}^2}, \\ \lambda_{m,h} &= \min_{E_m\in\mathcal{V}_{m,h}} \max_{\substack{v\in E_m \\ v_h\neq 0}} \frac{a(v_h,v_h)}{\|v_h\|_{L^2}^2}, \end{aligned}$$

et vérifient la propriété suivante :

$$\lambda_m \leq \lambda_{m,h}, \quad 1 \leq m \leq N_h.$$

De plus, l'égalité suivante nous permet d'obtenir un premier résultat de convergence.

$$\begin{aligned}
\lambda_{m,h} - \lambda_m &= a(u_{m,h}, u_{m,h}) - a(u_m, u_m) \\
&= a(u_{m,h} - u_m, u_{m,h} - u_m) + 2\, a(u_{m,h} - u_m, u_m) \\
&= a(u_m - u_{m,h}, u_m - u_{m,h}) - 2\, \lambda_m (u_m, u_m - u_{m,h})_{L^2} \\
&= a(u_m - u_{m,h}, u_m - u_{m,h}) - \lambda_m \Big[(u_m, u_m)_{L^2} + (u_{m,h}, u_{m,h})_{L^2} \\
&\qquad - 2(u_m, u_{m,h})_{L^2} \Big] \quad (\text{ en utilisant } \|u_m\|_{L^2} = \|u_{m,h}\|_{L^2} = 1) \\
&= a(u_m - u_{m,h}, u_m - u_{m,h}) - \lambda_m (u_m - u_{m,h}, u_m - u_{m,h})_{L^2}. \quad (3)
\end{aligned}$$

Avec la continuité de la forme bilinéaire a, ainsi que l'inégalité de Cauchy-Schwarz sur le second terme de droite, on a

$$|\lambda_{m,h} - \lambda_m| \leq C \|u_m - u_{m,h}\|^2_{H^1}. \quad (4)$$

Faisons maintenant l'hypothèse d'approximation suivante :
Il existe une constante $C > 0$ telle que pour tout $u \in H^1_0(\Omega) \cap H^\ell(\Omega)$,
$1 \leq \ell \leq k+1$ on ait :

$$\inf_{v_h \in X^k_h} \|u - v_h\|_{H^1} \leq C h^{\ell-1} \|u\|_{H^\ell}.$$

Les estimations d'erreur suivantes sont classiques, voir par exemple [2, 45, 50], néanmoins nous choisissons de donner ici une démonstration qui servira d'introduction au raisonnement utilisé pour obtenir les estimations d'erreur *a priori* dans le cas de problèmes aux valeurs propres non linéaires.

Lemme : *Si λ_m est une valeur propre simple, alors pour $h \leq h_0$ assez petit, $\lambda_{m,h}$ est une valeur propre simple et il existe une constante C positive, indépendante du sous-espace X^k_h, telle que l'on ait*

$$\|u_m - u_{m,h}\|_{H^1} \leq C h^{\ell-1} \|u\|_{H^\ell} \quad (5)$$

$$\|u_m - u_{m,h}\|_{L^2} \leq C h \|u_m - u_{m,h}\|_{H^1}. \quad (6)$$

Démonstration :
Seul le cas $m = 1$ sera traité. On commencera par montrer que

$$\|u_1 - u_{1,h}\|_{H^1} \underset{h \to 0}{\longrightarrow} 0. \quad (7)$$

Pour cela, on considère le problème de minimisation suivant :

$$I_{\text{lin}} = \inf\{E_{\text{lin}}(v), v \in H^1_0(\Omega), \|v\|_{L^2} = 1\},$$

où $E_{\text{lin}}(v) = \dfrac{1}{2} a(v, v)$. Ce problème admet une unique solution positive, que l'on notera u. Ainsi, en introduisant l'opérateur auto-adjoint, A_{lin}, tel que

$$\langle A_{\text{lin}} u, v \rangle = a(u, v) \qquad \forall v \in H^1_0(\Omega),$$

on obtient que la fonction u vérifie l'équation d'Euler-Lagrange suivante :

$$A_{\text{lin}}u = \lambda u$$

où $\lambda \in \mathbb{R}$ est le multiplicateur de Lagrange associé à la contrainte $\|u\|_{L^2} = 1$. Le problème aux valeurs propres, issu de ce problème de minimisation, n'est autre que (1). Il existe une base hilbertienne orthonormale de $L^2(\Omega)$ formée des vecteurs propres de (1), de façon à ce que tout $v \in L^2(\Omega)$ et de norme 1, puisse s'écrire sous la forme $v = \sum_m \alpha_m u_m$ avec $\sum_m \alpha_m^2 = 1$ et $a(v,v) = \sum_m \alpha_m^2 \lambda_m$. De ce fait, en remarquant λ_1 est la plus petite valeur propre, il apparaît que le couple (u_1, λ_1) est solution de cette équation d'Euler-Lagrange, ainsi que du problème de minimisation I_{lin}.
Par ailleurs, on peut montrer que λ_1, la plus petite valeur propre A_{lin}, est simple. Il en résulte que

$$\langle A_{\text{lin}}v, v\rangle F - \lambda_1 \int_\Omega v^2 \geq (\lambda_2 - \lambda_1)\left[\|v\|_{L^2}^2 - (u_1, v)_{L^2}^2\right] \qquad \forall v \in H_0^1(\Omega).$$

On rappelle que $\|u_1\|_{L^2(\Omega)} = \|u_{1,h}\|_{L^2(\Omega)} = 1$, de ce fait $|(u_1, u_{1,h})_{L^2(\Omega}| \leq 1$. Par conséquent, il apparaît

$$\langle A_{\text{lin}}(u_1 - u_{1,h}), (u_1 - u_{1,h})\rangle - \lambda_1 \int_\Omega (u_1 - u_{1,h})^2 \geq (\lambda_2 - \lambda_1)\left[1 - |(u_1, u_{1,h})_{L^2(\Omega)}|\right]$$

Ainsi, en choisissant $u_{1,h}$ tel que $(u_1, u_{1,h})_{L^2(\Omega)} \geq 0$, il découle

$$\begin{aligned}
a(u_1 - u_{1,h}, u_1 - u_{1,h}) &- \lambda_1 \|u_1 - u_{1,h}\|_{L^2(\Omega)}^2 \\
&= \langle A_{\text{lin}}(u_1 - u_{1,h}), (u_1 - u_{1,h})\rangle - \lambda_1 \|u_1 - u_{1,h}\|_{L^2(\Omega)}^2 \\
&\geq \frac{1}{2}(\lambda_2 - \lambda_1\left[\|u_1\|_{L^2(\Omega)} + \|u_{1,h}\|_{L^2(\Omega)}^2 - 2(u_1, u_{1,h})_{L^2(\Omega)}\right] \\
&\geq \frac{(\lambda_2 - \lambda_1)}{2}\|u_1 - u_{1,h}\|_{L^2(\Omega)}^2.
\end{aligned}$$

Soit θ tel que $0 < \theta \leq \dfrac{\lambda_2 - \lambda_1}{\lambda_2 + \lambda_1} < 1$. À partir de la ligne précédente, on obtient alors

$$\begin{aligned}
&a(u_1 - u_{1,h}, u_1 - u_{1,h}) - \lambda_1 \|u_1 - u_{1,h}\|_{L^2(\Omega)}^2 \\
&\quad = \theta\, a(u_1 - u_{1,h}, u_1 - u_{1,h}) + (1-\theta)\left[a(u_1 - u_{1,h}, u_1 - u_{1,h}) - \lambda_1 \|u_1 - u_{1,h}\|_{L^2(\Omega)}^2\right] \\
&\qquad -\theta\lambda_1 \|u_1 - u_{1,h}\|_{L^2(\Omega)}^2 \\
&\quad \geq \theta a(u_1 - u_{1,h}, u_1 - u_{1,h}) + \frac{(1-\theta)(\lambda_2 - \lambda_1)}{2}\|u_1 - u_{1,h}\|_{L^2(\Omega)}^2 - \theta\lambda_1 \|u_1 - u_{1,h}\|_{L^2(\Omega)}^2 \\
&\quad \geq \theta a(u_1 - u_{1,h}, u_1 - u_{1,h}) + \frac{1}{2}\left[\lambda_2 - \lambda_1 - \theta(\lambda_2 + \lambda_1)\right]\|u_1 - u_{1,h}\|_{L^2(\Omega)}^2.
\end{aligned}$$

Ceci implique, en utilisant la coercivité de la forme bilinéaire a et en choisissant θ assez petit, qu'il existe une constante C positive telle que

$$a(u_1 - u_{1,h}, u_1 - u_{1,h}) - \lambda_1 (u_1 - u_{1,h}, u_1 - u_{1,h})_{L^2} \geq C\|u_1 - u_{1,h}\|_{H^1(\Omega)}^2. \quad (8)$$

En combinant ceci avec l'égalité (3), on a

$$\begin{aligned}\frac{C}{2}\|u_1-u_{1,h}\|^2_{H^1(\Omega)} &\leq \frac{1}{2}a(u_1-u_{1,h},u_1-u_{1,h})-\frac{\lambda_1}{2}(u_1-u_{1,h},u_1-u_{1,h})_{L^2}\\ &= \frac{1}{2}(\lambda_{1,h}-\lambda_1)=E_{\text{lin}}(u_{1,h})-E_{\text{lin}}(u_1).\end{aligned}$$

Puisque $u_{1,h}$ est solution du problème aux valeurs propres (2), elle est également le minimiseur de E_{lin} sur l'espace X_h^k. On a ainsi pour tout $v_h \in X_h^k$, $E_{\text{lin}}(u_{1,h}) \leq E_{\text{lin}}(v_h)$. Soit $x_h \in X_h^k$ telle que

$$\|u_1-x_h\|_{H^1(\Omega)} = \inf_{v_h\in X_h^k}\|u_1-v_h\|_{H^1(\Omega)} \text{ et } \|u_1-x_h\|_{H^1(\Omega)} \underset{h\to 0}{\longrightarrow} 0.$$

On obtient finalement

$$\begin{aligned}\frac{C}{2}\|u_1-u_{1,h}\|^2_{H^1(\Omega)} &\leq E_{\text{lin}}(x_h)-E_{\text{lin}}(u_1)\\ &= \frac{1}{2}a(v_h,v_h)-\frac{1}{2}a(u_1,u_1)=\frac{1}{2}a(x_h-u_1,v_h+u_1)\\ &\leq C\|x_h-u_1\|_{H^1(\Omega)}\|x_h+u_1\|_{H^1(\Omega)} \underset{h\to 0}{\longrightarrow} 0.\end{aligned}$$

Revenons à la démonstration de l'estimation (5). En utilisant (8), pour tout $x_h \in X_h^k$, on a

$$\begin{aligned}C\|u_1-u_{1,h}\|^2_{H^1(\Omega)} &\leq a(u_1-u_{1,h},u_1-u_{1,h})-\lambda_1\|u_1-u_{1,h}\|^2_{L^2(\Omega)}\\ &\leq a(u_1-x_h,u_1-u_{1,h})-\lambda_1\int_\Omega (u_1-x_h)(u_1-u_{1,h})\\ &\quad +a(x_h-u_{1,h},u_1-u_{1,h})-\lambda_1\int_\Omega (x_h-u_{1,h})(u_1-u_{1,h})\\ &\leq a(u_1-x_h,u_1-u_{1,h})-\lambda_1\int_\Omega (u_1-x_h)(u_1-u_{1,h})\\ &\quad +\lambda_1\int_\Omega (x_h-u_{1,h})u_1-\lambda_{1,h}\int_\Omega (x_h-u_{1,h})u_{1,h}\\ &\quad -\lambda_1\int_\Omega (x_h-u_{1,h})(u_1-u_{1,h})\\ &\leq a(u_1-x_h,u_1-u_{1,h})-\lambda_1\int_\Omega (u_1-x_h)(u_1-u_{1,h})\\ &\quad +(\lambda_1-\lambda_{1,h})\int_\Omega (x_h-u_{1,h})u_{1,h}\\ &\leq C\|u_1-x_h\|_{H^1(\Omega)}\|u_1-u_{1,h}\|_{H^1(\Omega)}+\ |\lambda_1-\lambda_{1,h}|\,\|u_1-x_h\|_{L^2(\Omega)}.\end{aligned}$$

Ainsi en utilisant (4) dans cette dernière ligne, il vient

$$\begin{aligned}\|u_1-u_{1,h}\|^2_{H^1(\Omega)} &\leq C\|u_1-u_{1,h}\|_{H^1(\Omega)}\Big[\|u_1-x_h\|_{H^1(\Omega)}+\|u_1-u_{1,h}\|_{H^1(\Omega)}\|u_1-x_h\|_{L^2(\Omega)}\Big]\\ &\leq C\|u_1-u_{1,h}\|_{H^1(\Omega)}\|u_1-x_h\|_{H^1(\Omega)}\Big[1+\|u_1-u_{1,h}\|_{H^1(\Omega)}\Big].\end{aligned}$$

D'après (7), le terme $\|u_1 - u_{1,h}\|_{H^1(\Omega)}$ est petit, de ce fait on trouve

$$\|u_1 - u_{1,h}\|_{H^1(\Omega)} \leq C\|u_1 - x_h\|_{H^1(\Omega)} = C \inf_{v_h \in X_h^k} \|u_1 - v_h\|_{H^1(\Omega)},$$

et finalement, si $u \in H^1(\Omega) \cap H^\ell(\Omega)$, $1 \leq \ell \leq k$, on a

$$\|u_1 - u_{1,h}\|_{H^1} \leq Ch^{\ell-1}\|u\|_{H^\ell}.$$

Il reste à montrer l'estimation (6), la démonstration qui suit diffère légèrement de celles qui existent dans la littérature, mais elle pourra facilement être adaptée pour fonctionner dans le cas de problème aux valeurs propres non linéaire. Pour cela, on note $u^\perp = \{v \in H_0^1(\Omega), \int_\Omega uv = 0\}$, le sous espace de $H_0^1(\Omega)$, et on considère le problème adjoint suivant :

Trouver $\psi \in u^\perp$ tel que pour tout $v \in u^\perp$, alors

$$a(\psi, v) - \lambda(\psi, v) = \int_\Omega (u_1 - u_{1,h})v. \tag{9}$$

La forme bilinéaire $(v, w) \mapsto a(w, v) - \lambda(w, v)$ étant coercive, continue et symétrique sur $u^\perp$, le théorème de Lax-Milgram nous assure l'existence et l'unicité de la solution ψ du problème (9). De plus elle vérifie les hypothèses de régularité et de continuité suivantes :

$$\psi \in H_0^1(\Omega) \cap H^2(\Omega) \tag{10}$$

$$\|\psi\|_{H^2(\Omega)} \leq C\|u_1 - u_{1,h}\|_{L^2(\Omega)}. \tag{11}$$

En particulier, on a

$$\inf_{v_h \in X_h^k} \|\psi - v_h\|_{H^1(\Omega)} \leq Ch\|u_1 - u_{1,h}\|_{L^2(\Omega)}. \tag{12}$$

Soit $u_1^* \in H_0^1(\Omega)$, définie par

$$u_1^* = u_{1,h} + (1 - \int_\Omega u_1\, u_{1,h})u_1,$$

de sorte que $u_1^* - u_1 \in u^\perp$, on a également

$$u_1^* - u_{1,h} = \frac{1}{2}u_1\|u_1 - u_{1,h}\|^2_{L^2(\Omega)}.$$

Alors, on remarque

$$\begin{aligned}
\|u_1-u_{1,h}\|^2_{L^2(\Omega)} &= \int_\Omega (u_1-u_{1,h})(u_1-u_1^*)+\int_\Omega ((u_1-u_{1,h})(u_1^*-u_{1,h}) \\
&= \int_\Omega (u_1-u_{1,h})(u_1-u_1^*)+\frac{1}{4}\|u_1-u_{1,h}\|^4_{L^2(\Omega)} \\
&= a(\psi,u_1-u_{1,h})-\lambda_1\int_\Omega \psi(u_1-u_{1,h}) \\
&= a(\psi-\psi_h,u_1-u_{1,h})-\lambda_1\int_\Omega (\psi-\psi_h)(u_1-u_{1,h}) \\
&\quad +a(\psi_h,u_1-u_{1,h})-\lambda_1\int_\Omega \psi_h(u_1-u_{1,h}),\quad \forall\psi_h\in X_h^k \\
&= a(\psi-\psi_h,u_1-u_{1,h})-\lambda_1\int_\Omega (\psi-\psi_h)(u_1-u_{1,h}) \\
&\quad +(\lambda_1-\lambda_{1,h})\left[\int_\Omega (\psi_h-\psi)u_{1,h}+\int_\Omega \psi u_{1,h}\right],\quad \forall\psi_h\in X_h^k.
\end{aligned}$$

Par conséquent, en utilisant (4), on obtient, pour tout $\psi_h \in X_h^k$

$$\begin{aligned}
\|u_1-u_{1,h}\|^2_{L^2(\Omega)} &\le C\|u_1-u_{1,h}\|_{H^1(\Omega)}\|\psi-\psi_h\|_{H^1(\Omega)} \\
&\quad +\|u_1-u_{1,h}\|^2_{H^1(\Omega)}\ \|u_1\|_{L^2(\Omega)}\left[\|\psi-\psi_h\|_{L^2(\Omega)}+\|\psi\|_{L^2(\Omega)}\right] \\
&\le C\|u_1-u_{1,h}\|_{H^1(\Omega)}\inf_{\psi_h\in X_h^k}\|\psi-\psi_h\|_{H^1(\Omega)} \\
&\quad +\|u_1-u_{1,h}\|^2_{H^1(\Omega)}\left[\inf_{\psi_h\in X_h^k}\|\psi-\psi_h\|_{L^2(\Omega)}+\|\psi\|_{H^2(\Omega)}\right] \\
&\le C\|u_1-u_{1,h}\|_{H^1(\Omega)}\|u_1-u_{1,h}\|_{L^2(\Omega)}\left[h+\|u_1-u_{1,h}\|_{H^1(\Omega)}\right] \\
&\le C\|u_1-u_{1,h}\|_{H^1(\Omega)}\|u_1-u_{1,h}\|_{L^2(\Omega)}\left[h+h^\ell\|u_1\|_{H^{\ell+1}(\Omega)}\right]\quad 1\le\ell\le k. \\
&\le Ch\|u_1-u_{1,h}\|_{H^1(\Omega)}\|u_1-u_{1,h}\|_{L^2(\Omega)}.
\end{aligned}$$

On retrouve ainsi (6), ce qui conclut la démonstration de ce lemme.
□

À la recherche du fondamental

Il existe plusieurs modèles mathématiques issus des sciences physiques et de l'ingénierie, dont la résolution demande une recherche de valeurs et vecteurs propres de problèmes non linéaires, comme le calcul de modes de vibration en mécanique des solides non linéaires. On trouve aussi des exemples en chimie quantique, où les modèles dit *ab initio* dérivent directement de l'équation de Schrödinger [27] :

- les équations de Gross-Pitaevskii qui décrivent les états stationnaires du condensat de Bose-Einsten,
- les modèles d'Hartree-Fock et Kohn-Sham.

Ces deux derniers ont pour but de déterminer l'état fondamental, c'est-à-dire l'état de plus basse énergie d'un système moléculaire.
Le modèle de Kohn- Sham est très populaire en physique du solide, mais aussi en chimie moléculaire. Il repose sur la Théorie de la fonctionnelle de densité (DFT, Density Functional Theory [23,24,41,53]). Le principal intérêt de cette méthode réside dans le fait qu'elle permet de modéliser des systèmes relativement étendus (molécules de taille importante ou des solides) avec une bonne précision. La description quantique non-relativiste d'un système moléculaire ou cristallin est basée sur l'équation de Schrödinger [27] suivante (qui sera simplifiée par diverses approximations pour faciliter sa résolution) :

$$H\Psi(\overrightarrow{R_j}, \overrightarrow{r_i}) = i\hbar\frac{\partial}{\partial t}\Psi(\overrightarrow{R_j}, \overrightarrow{r_i})),$$

où H est le hamiltonien du système. Cette équation peut être ramenée à un cas stationnaire, qui prend la forme d'un problème aux valeurs propres

$$H\psi(\overrightarrow{R_j}, \overrightarrow{r_i}) = E\psi(\overrightarrow{R_j}, \overrightarrow{r_i}),$$

où $\psi(\overrightarrow{R_j}, \overrightarrow{r_i})$ est la fonction d'onde qui décrit le comportement des électrons, $\overrightarrow{R_j}$ et $\overrightarrow{r_i}$ représentent les coordonnées des noyaux et des électrons, E correspond quant à lui à l'énergie du système. Chaque valeur propre E correspond à un niveau d'énergie associé à un état du système décrit par la fonction d'onde $\psi(\overrightarrow{R_j}, \overrightarrow{r_i})$. L'état de plus faible énergie est le plus stable, chercher celui-ci revient à résoudre un problème de minimisation. Pour un système moléculaire composé de M noyaux et de N électrons. La complexité de ce problème est telle qu'il ne peut être résolu sans simplification supplémentaire :

- 1^{er} étape d'approximation : l'approximation de Born-Oppenheimer,
- $2^{\text{ème}}$ étape d'approximation : les méthodes de type Fonctionnelle de la densité ou Hartree-Fock (celui-ci ne sera pas traité dans cette thèse),
- $3^{\text{ème}}$ étape d'approximation : les méthodes de discrétisations et de résolutions numériques.

Le premier niveau d'approximation est basé sur l'approximation de Born-Oppenheimer, qui permet de traiter séparément les électrons et les noyaux d'un système. Celle-ci s'appuie sur la différence de masse entre ces deux familles de particules. Ainsi, on peut découpler le mouvement des noyaux de celui des électrons. On fixe alors la position des noyaux. Ils deviennent des paramètres et les degrés de liberté nucléaires apparaissent uniquement dans le potentiel moyen W. La position d'équilibre la plus stable du système est donc obtenue minimisant l'énergie potentiel W.
Cette énergie comprend deux termes :

1. le terme $\sum_{1\leq k\leq l\leq M} \frac{z_k z_l}{|\overrightarrow{R_k}-\overrightarrow{R_l}|}$ qui décrit la répulsion internucléaire (où z_i représente la charge du noyau i),
2. le terme qui correspond au potentiel effectif ressenti par les noyaux, dû à la présence du nuage électronique. La valeur de ce potentiel en un point est obtenue en cherchant le fondamental du hamiltonien électronique H_e sur l'espace des fonctions d'ondes, que l'on appellera problème electronique.

En raison de la taille des fonctions d'ondes, la résolution numérique de ce problème de minimisation, telle quelle, n'est possible que pour des systèmes ne contenant qu'un ou deux atomes. Il existe ainsi un second niveau d'approximation découpé en deux classes : la méthode d'Hartree - Fock et celle issue de la théorie de la Fonctionnelle de la densité.
La méthode de Hartree-Fock ([21, 48]) est une approximation variationnelle du problème électronique consistant à restreindre l'ensemble de minimisation aux seules fonctions d'onde ψ_e, qui s'écrivent comme un déterminant de Slater de N fonctions d'onde monoélectroniques orthonormées ϕ_i appelées *orbitales moléculaires*

$$\psi_e(\overrightarrow{r_1},\cdots,\overrightarrow{r_N}) = \frac{1}{\sqrt{(N!)}}\det(\phi_i(\overrightarrow{r_j})).$$

Soit

$$\mathcal{W}_N = \{\Phi=(\phi_i)_{1\leq i\leq N},\ \phi_i\in H^1(\mathbb{R}^3),\ \text{ tel que } \int_{\mathbb{R}^3}\phi_i\phi_j=\delta_{i,j},\ 1\leq i,j\leq N\}$$

l'ensemble des configurations de N orbitales moléculaires.
On note par $\rho_\Phi(x)=\sum_{i=1}^N|\phi(x)|^2$ la densité électronique, $\tau_\Phi(x,x')=\sum_{i=1}^N\phi_i(x)\phi_i(x')$ la matrice densité d'ordre 1 et $V(x)=-\sum_{k=1}^M\frac{z_k}{|x-\overrightarrow{R_k}|}$, le potentiel créé par les noyaux et subis par les électrons du système. Le problème d'Hartree-Fock s'écrit sous la forme

$$\inf\{E^{HF}(\Phi),\Phi\in\mathcal{W}_N\},$$

avec

$$\begin{aligned} E^{HF}(\Phi) &= \sum_{i=1}^N\frac{1}{2}\int_{\mathbb{R}^3}|\nabla\phi_i|^2+\int_{\mathbb{R}^3}V\rho_\Phi+\frac{1}{2}\int_{\mathbb{R}^3}\int_{\mathbb{R}^3}\frac{\rho_\Phi(x)\rho_\Phi(x')}{|x-x'|}dxdx' \\ &\quad -\frac{1}{2}\int_{\mathbb{R}^3}\int_{\mathbb{R}^3}\frac{|\tau_\Phi(x,x')|^2}{|x-x'|}dxdx'. \end{aligned}$$

Le premier terme représente l'énergie cinétique de la fonction d'onde, le troisième terme correspond à la répulsion coulombienne, quant au dernier il résulte de l'antisymétrie de la fonction d'onde, et est appelé *terme d'échange*. Tout

minimum du problème de Hartree-Fock vérifie les équations d'Euler-Lagrange du problème suivant, à savoir

$$-\frac{1}{2}\Delta\phi_i + V\phi_i + \left(\rho_\Phi \star \frac{1}{|x|}\right)\phi_i - \int_{\mathbb{R}^3} \frac{|\tau_\Phi(x,y)|^2}{|x-y|}dxdy = \epsilon_i\phi_i, \qquad \forall 1 \leq i \leq N,$$

où ϵ_i est le multiplicateurs de Lagrange associé à la contrainte $\int_{\mathbb{R}^3} \phi_i\phi_j = \delta_{i,j}$. Ce modèle très non linéaire reste très complexe, et très coûteux en terme de calcul. L'existence d'un état fondamental est connue ([26,30–32]), en revanche l'unicité de celui-ci reste un problème ouvert. Par ailleurs, des résultats sur l'analyse numérique de ce modèle, en particulier sur les estimations *a posteriori*, ont été établis dans [37].

La seconde classe de modèle est celle issue de la Théorie de la Fonctionnelle de la densité. Celle-ci est très populaire en physique du solide et gagne en succès en chimie moléculaire. Elle consiste à utiliser la densité électronique comme variable principale pour caractériser le système. Ainsi, contrairement aux modèles de type Hartree-Fock, où la variable est une fonction d'onde multi-électronique, avec $3 \times N$ degrés de liberté, elle n'en a ici plus que trois. Pour déterminer l'énergie et la densité électronique fondamentale, il suffit de résoudre directement un problème de minimisation de la forme

$$\inf\left\{F(\rho) + \int_{\mathbb{R}^3} \rho V,\ \rho \in L^1(\mathbb{R}^3),\ \rho \geq 0,\ \int_{\mathbb{R}^3} \rho = N\right\},$$

où F est une fonctionnelle *universelle*, c'est à dire qu'elle ne dépend pas du potentiel V créé par les noyaux. Avant même qu'une justification théorique soit apportée par Hohenberg et Kohn en 1964 [23], il existait déjà des modèles utilisant ce type de formalisme. Ce sont les modèles dit de Thomas- Fermi ([14, 16,51]), qui sont apparus dans les années 30. La fonctionnelle F ne pouvant être exprimée explicitement, elle a été approchée à l'aide de modèles empiriques de la physique statistique des gaz homogènes d'électrons. Dans le premier modèle proposé, celui de Thomas-Fermi, la fonctionnelle $F(\rho)$ est remplacée par :

$$\mathcal{F}_{TF}(\rho) = C_{\mathrm{TF}} \int_{\mathbb{R}^3} \rho^{5/3} + \frac{1}{2}\int_{\mathbb{R}^3}\int_{\mathbb{R}^3} \frac{\rho_\Phi(x)\rho_\Phi(y)}{|x-y|}dxdy,$$

avec $C_{\mathrm{TF}} = \frac{3^{5/3}\pi^{4/3}}{10}$. Le premier terme correspond à l'énergie cinétique d'un gaz homogène d'électrons et peut être déterminé simplement. Le second terme, dit d'échange, décrit l'interaction coulombienne. Ce modèle a été perfectionné par von Weizsäcker en améliorant le terme d'énergie cinétique. La fonctionnelle F est alors approchée par

$$\mathcal{F}_{TFW}(\rho) = C_W \int_{\mathbb{R}^3} |\nabla\sqrt{\rho}|^2 + \mathcal{F}_{TF}(\rho),$$

avec $C_W = 0.093$. Plus tard, le terme d'échange a été corrigé par Dirac pour mieux décrire l'intéraction entre les électrons, c'est ainsi que le modèle Thomas-Fermi-Dirac-von Weizsäcker a été introduit :

$$\mathcal{F}_{TFDW}(\rho) = \mathcal{F}_{TFW}(\rho) - C_D \int_{\mathbb{R}^3} \rho^{4/3},$$

avec $C_D = \frac{3}{4}\left(\frac{3}{\pi}\right)^{1/3}$.

Dans le modèle de Kohn-Sham, la fonctionnelle F est décomposée en trois contributions $F(\rho) = T_{KS}(\rho) + J(\rho) + E_{xc}(\rho)$, où T_{KS} est l'énergie cinétique d'un système non interagissant, J est l'énergie coulombienne et E_{xc} est l'énergie dite d'échange-correlation, avec

$$\begin{aligned} T_{KS} &= \inf\left\{\frac{1}{2}\sum_{i=1}^{N}\int_{\mathbb{R}^3}|\nabla\phi_i|^2,\ \Phi=\{\phi_i\}\in\mathcal{W}_N,\ \rho_\Phi=\rho\right\} \\ J(\rho) &= \frac{1}{2}\int_{\mathbb{R}^3}\int_{\mathbb{R}^3}\frac{\rho_\Phi(x)\rho_\Phi(y)}{|x-y|}dxdy \\ E_{xc} &= F(\rho) - T_{KS} - J(\rho). \end{aligned}$$

La méthode de Kohn-Sham est formellement exacte mais la fonctionnelle E_{xc} est inconnue. De ce fait, la validité de ce modèle dépend exclusivement de la qualité de la fonctionnelle d'échange-corrélation approchée. Dans le cas de Kohn-Sham, on ne cherche pas à approcher directement la fonctionnelle F, mais seulement une partie, ce qui donne de meilleurs résultats . En effet, bien qu'ils reproduisent correctement un certain nombres de phénomènes, les modèles de type Thomas-Fermi ne sont guère utilisés en chimie. Ils restent toutefois intéressants d'un point de vue mathémarique, car malgré leur simplicité par rapport aux modèles de type Hartree-Fock ou Kohn-Sham, les difficultés restent semblables (non linéarité des équations, présence de potentiels coulombiens et de fonctionnelles non locales). De nombreuses études ont été réalisés sur ces modèles, mais il existe peu de travaux sur leur analyse numérique [28,59,60]. La première partie de cette thèse est essentiellement consacrée à l'amélioration des estimations *a priori* déjà existantes pour le modèle de Thomas- Fermi.

Quelques résultats d'estimations *a priori* dans le cas de problèmes aux valeurs propres non linéaires

Les problèmes aux valeurs qui nous intéressent sont de la forme suivante :

Trouver $u \in X$ et $\lambda \in \mathbb{R}$ tels que

$$-\ \mathrm{div}(A\cdot\nabla u) + Vu + G'(u^2)u = \lambda u \ \text{ sur } \Omega, \tag{13}$$

avec Ω dans $\mathbb{R}^d$, $d = 1, 2, 3$. Lorsque les conditions aux bords sont périodiques, le domaine Ω est le cube $(0, 2\pi)^d$ et X désignera l'espace $H^1_\#(\Omega)$ défini par

$$H^1_\#(\Omega) = \{v \in H^1(\Omega),\ \text{v est } 2\pi\text{-périodique sur } \Omega\}.$$

Dans le cas contraire, Ω sera simplement un domaine borné à frontière régulière, et X sera l'espace $H^1_0(\Omega)$. En plus de ces conditions au bord, il est courant d'imposer une contrainte de normalisation sur les fonctions u, c'est-à-dire

$\|u\|_{L^2(\Omega)} = 1$.
Ce problème aux valeurs propres est issu du problème de minimisation suivant :

$$I = \inf\{E(w),\, w \in X, \|w\|_{L^2(\Omega)} = 1\},$$

avec

$$E(w) = \frac{1}{2}\int_\Omega (A\nabla w)\cdot(\nabla w) + \frac{1}{2}\int_\Omega V w^2 + \int_\Omega G(w^2),$$

qui a exactement deux solutions u et $-u$. On notera u la solution positive. Celle-ci vérifie l'équation de Euler-Lagrange suivante

$$\forall v \in X, \quad \int_\Omega (A\nabla u)\cdot(\nabla v) + \int_\Omega V uv + \int_\Omega G'(u^2)uv - \lambda\int_\Omega uv = 0,$$

où $\lambda \in \mathbb{R}$ est le multiplicateur de Lagrange associé à la contrainte $\|u\|_{L^2(\Omega)} = 1$. Cette équation d'Euler-Lagrange, avec la contrainte de normalisation, prend la forme d'un problème aux valeurs propres non linéaire qui n'est autre que (13). Par ailleurs, λ, la plus petite valeur propre de (13), est simple, et u est la fonction propre associée à λ.
Soit X_δ une famille de sous espace de dimension finie de X telle que pour tout $v \in X$

$$\inf_{v_\delta \in X_\delta} \{\|v - v_\delta\|_{H^1(\Omega)}\} \underset{\delta\to\delta_\infty}{\longrightarrow} 0.$$

On définit le problème de minimisation discret suivant

$$I_\delta = \inf\{E(w_\delta),\, w_\delta \in X_\delta, \|w_\delta\|_{L^2(\Omega)} = 1\}. \tag{14}$$

Celui-ci admet exactement deux minimiseurs u_δ et $-u_\delta$ qui vérifient le problème aux valeurs propres

$$\forall v_\delta \in X_\delta, \quad \int_\Omega (A\nabla u_\delta)\cdot(\nabla v_\delta) + \int_\Omega V u_\delta v_\delta + \int_\Omega G'(u_\delta^2)u_\delta v_\delta - \lambda_\delta\int_\Omega u_\delta v_\delta = 0.$$

Lemme : (voir par exemple [59, 60])
Il existe un $\delta_0 > 0$ et C, γ, et $M \in \mathbb{R}_+$ tels que pour tout $0 < \delta < \delta_0$ on ait

$$\|u - u_\delta\|_{H^1(\Omega} \underset{\delta\to\delta_\infty}{\longrightarrow} 0$$

$$\frac{\gamma}{2}\|u - u_\delta\|^2_{H^1(\Omega} \le E(u_\delta) - E(u) \le \frac{M}{2}\|u - u_\delta\|^2_{H^1(\Omega)},$$

$$\begin{aligned} |\lambda_\delta - \lambda| &\le C\left(\|u_\delta - u\|^2_{H^1(\Omega} + \int_\Omega (u_\delta - u)u_\delta^2 \frac{G'(u_\delta^2) - G'(u^2)}{u_\delta - u}\right) \\ |\lambda_\delta - \lambda| &\le C(\|u_\delta - u\|^2_{H^1(\Omega} + \|u_\delta - u\|_{L^2(\Omega)}. \end{aligned} \tag{15}$$

Cette estimation sur les valeurs propres est très décevante, en particulier lorsqu'on la compare à celle obtenue dans le cas linéaire. Pour améliorer celle-ci, il faudra traiter différemment l'intégrale $\displaystyle\int_\Omega (u_\delta - u)u_\delta^2 \frac{G'(u_\delta^2) - G'(u^2)}{u_\delta - u}$ de façon à

faire ressortir une norme négative.

Deux types de discrétisations ont été analysées : une méthode spectrale (ondes planes) et la méthode des éléments finis.
Dans le cas d'un problème aux valeurs propres non linéaires, ayant des conditions aux limites périodiques, il est naturel d'utiliser une base d'ondes planes pour discrétiser l'espace X.
On note $e_k(x) = \dfrac{e^{ik.x}}{2\pi^{d/2}}$, pour $k \in \mathbb{Z}^d$, de façon à ce que pour tout $v \in L^2_{\#}(\Omega)$, on ait $v(x) = \sum_{k\in\mathbb{Z}^d} \widehat{v}_k e_k(x\ )$, où $\widehat{v}_k$ désigne le $k^{\text{ème}}$ coefficient de Fourier de v.
On choisit ainsi $X_\delta = X_N = \text{Span}\{e_k,\ |k|_\infty \leq N\}$, de sorte que, pour tout $v \in H^s_{\#}(\Omega)$, sa meilleure approximation dans $H^r_{\#}(\Omega)$ pour tout $r \leq s$ soit

$$\Pi_N v = \sum_{k\in\mathbb{Z}^d, |k|_\infty \leq N} \widehat{v}_k e_k,$$

et

$$\forall v \in H^s_{\#}(\Omega) \qquad \|\ v - \Pi_N v\|_{H^r(\Omega)} \leq \frac{1}{N^{s-r}} \|v\|_{H^s_{\#}(\Omega)}.$$

Par ailleurs on notera u_N la solution discrète u_δ et l'on supposera que $V \in\ H^\sigma_{\#}(\Omega)$, pour $\sigma > d/2$.

Théorème (Ondes Planes) :
Pour tout $N \in \mathbb{N}$, on note u_N le minimiseur de (14), tel que $(u, u_N)_{L^2(\Omega)} \geq 0$. Alors, pour N assez grand, u_N est unique et vérifie les estimations suivantes

$$\|u - u_N\|_{H^s(\Omega)} \leq \frac{C}{N^{\sigma+2-s}} \|u\|_{H^{\sigma+2}(\Omega)} \qquad -\sigma \leq s < \sigma + 2,$$

$$|\lambda_N - \lambda| \leq C \frac{C}{N^{2(\sigma+1)}} \|u\|_{H^{\sigma+2}(\Omega)}.$$

Par ailleurs, lorsque les conditions aux bord ne sont pas périodiques, on choisit X^k_h, l'espace de type éléments finis, pour approcher X. De plus, on posera $u_h = u_\delta$.

Théorème (Eléments Finis) :
Pour tout h, on note u_h le minimiseur de (14), tel que $(u, u_h)_{L^2(\Omega)} \geq 0$. Alors pour h assez petit, u_h est unique et il vérifie les estimations suivantes :

- *Si $V \in L^2(\Omega)$, et si X_h est un espace de type éléments finis $\mathbb{P}_1$ on a*

$$\begin{aligned} \|u - u_h\|_{H^1(\Omega)} &\leq Ch\|u\|_{H^2(\Omega)} \\ \|u - u_h\|_{L^2(\Omega)} &\leq Ch^2\|u\|_{H^2(\Omega)} \\ |\lambda_h - \lambda| &\leq Ch^2\|u\|_{H^2(\Omega)}. \end{aligned}$$

– *Si* $V \in H^1(\Omega)$*, et que* X_h *est un espace de type éléments finis* $\mathbb{P}_2$ *on a*

$$\begin{aligned} \|u - u_h\|_{H^1(\Omega)} &\leq Ch^2\|u\|_{H^3(\Omega)} \\ \|u - u_h\|_{L^2(\Omega)} &\leq Ch^3\|u\|_{H^3(\Omega)} \\ \|u - u_h\|_{H^{-1}(\Omega)} &\leq Ch^4\|u\|_{H^3(\Omega)} \\ |\lambda_h - \lambda| &\leq Ch^4\|u\|_{H^3(\Omega)}. \end{aligned}$$

Grâce aux estimations d'erreur obtenues en norme négative, on retrouve des résultats du même ordre de convergence que ceux du cas linéaire. Toutefois, il faut tenir compte de l'effet de l'intégration numérique. En effet, le choix du nombre de points d'intégration N_g est primordial, car une sous-intégration pourrait gravement détériorer ces estimations (en particulier pour les valeurs propres).
La figure suivante 1 illustre ce phénomène lors de l'approximation par des ondes planes, pour différentes valeurs de N et de N_g, du problème aux valeurs propres non linéaire de dimension 1 :

Trouver $(u, \lambda) \in X \times \mathbb{R}$, tels que $\|u\|_{L^2} = 1$ et

$$-\Delta u + Vu + u^3 = \lambda u,$$

avec $V(x) = sin(|x - \pi|/2)$, $4 \leq N \leq 30$, $N_g = 2^p$ et $7 \leq p \leq 15$.

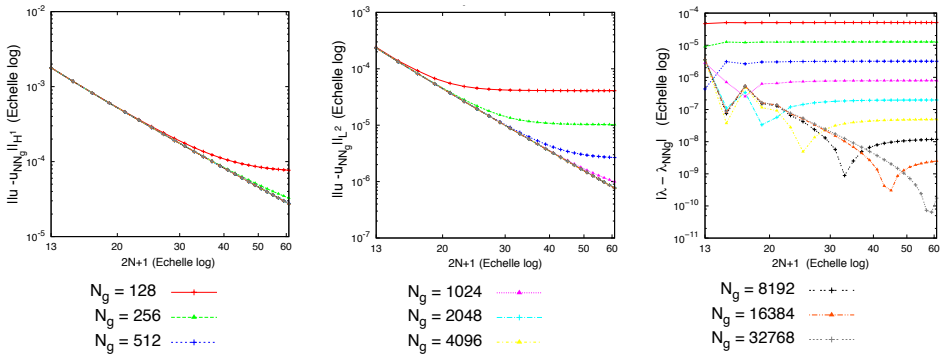

FIG. 1 – Effet de l'intégration numérique sur les taux de convergence

Méthodes de sous-grilles pour la résolution de problèmes aux valeurs propre non linéaires

Les estimations d'erreurs vues précédemment sont nécessaires à l'adaptation des *schémas à deux grilles,* pour la résolution de problèmes aux valeurs propres non linéaires. En effet, cette technique repose sur le fait que la contribution de u_δ à l'erreur mesurée en norme $L^2(\Omega)$ a un ordre plus élevé que si elle était mesurée en norme $H^1(\Omega)$.
L'utilisation de ces *schémas à deux grilles d'éléments finis*, pour la résolution

de problèmes aux valeurs propres, a été introduite par Xu et Zhou [57, 58] et plus particulièrement, pour les équations de Kohn-Sham. Celles-ci sont de la forme

$$[-\Delta + V + V_s(\rho)]\phi_i = \epsilon_i \phi_i.$$

La particularité de ce problème est que le terme V_s est non linéaire. Celui-ci provient de la dérivation de l'énergie coulombienne et de celle de l'échange-correlation. En effet, ce potentiel V_s dépend de la densité électronique ρ (qui n'est autre que la somme des $|\phi_i|^2$). Ce type d'équation doit être résolue à l'aide d'algorithmes itératifs dit *self consistant* (SCF) (*cf.* schéma ci-dessous).

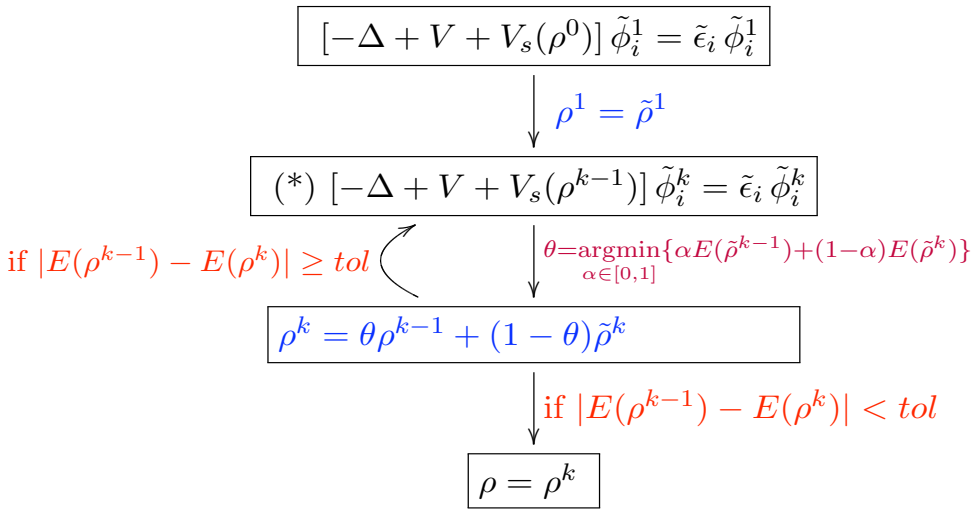

FIG. 2 – Schéma de résolution d'un problème aux valeurs propres non linéaire à l'aide d'un algorithme de type SCF

Ainsi, à chaque étape (*), il faut résoudre un problème aux valeurs propres linéaire. Dans les schémas à deux grilles existant [12], cette étape est remplacée par la résolution d'un problème linéarisé avec un second membre. On a par exemple ces différentes étapes :

1. Résoudre le problème aux valeurs propres linéaires sur un espace de discrétisation grossier noté S_H :

   Trouver $(u_H, \lambda_H) \in S_H \times \mathbb{R}$, tels que
   $\|u_H\|_{L^2(\Omega)} = 1$ et

   $$\int_\Omega \nabla u_H \nabla v + \int_\Omega (V + V_s) u_H v = \lambda_H \int_\Omega u_h v \quad \forall v \in S_H.$$

2. Résoudre le problème linéaire avec second membre sur un espace de discrétisation fin S^h :

Trouver $u_h \in S_h$ telle que

$$\int_\Omega \nabla u_h \nabla v + \int_\Omega (V + V_s) u_h v = \lambda_H \int_\Omega u_h v \quad \forall v \in S_h.$$

3. Poser $\lambda_h = \left(\int_\Omega |\nabla u_h|^2 + \int_\Omega (V + V_s) u_h^2 \right) \Big/ \left(\|u_h\|^2_{L^2(\Omega)} \right)$

Bien que cette méthode donne de bon résultat (car la résolution d'un problème linéaire avec second membre est plus rapide et moins complexe que celle d'un problème aux valeurs propres linéaire), la partie itérative de cet algorithme reste néanmoins faite sur la grille la plus fine.

C'est pourquoi les schémas présentés ici sont tels que, dans l'espace de discrétisation fin, ne sont résolus que des problèmes linéaires. Pour cela, il faut préalablement calculer la solution du problème aux valeurs propres non linéaire sur un espace grossier. Cette solution sera ensuite utilisée pour résoudre un problème aux valeurs propres linéarisé, ou même un problème linéarisé avec second membre, de la façon suivante :

1. Sur une grille grossière

Résolution d'un problème aux valeurs propres non linéaire sur un espace grossier X_H^1
$a_{\text{lin}}(u_H, v) + \int_\Omega G'(u_H^2) u_H v = \lambda_H \int_\Omega u_H v, \quad \forall v \in X_H^1$

2. Sur une grille fine

Problème 1	Problème 2	Problème 3
Résolution d'un problème aux valeurs propres linéarisé sur un espace fin X_h^1	Résolution d'un problème linéarisé avec second membre sur un espace fin X_h^1	Résolution d'un problème linéarisé avec second membre sur un espace fin X_h^1
$a_{\text{lin}}(u_h^H, v) + \int_\Omega G'(u_H^2) u_h^H v = \lambda_h^H \int_\Omega u_h^H v \quad \forall v \in X_h^1$	$a_{\text{lin}}(\tilde{u}_h^H, v) + \int_\Omega G'(u_H^2) \tilde{u}_h^H v = \lambda_H \int_\Omega u_H v \quad \forall v \in X_h^1$	$a_{\text{lin}}(\overline{u}_h^H, v) = -\int_\Omega G'(u_H^2) u_H v + \lambda_H \int_\Omega u_H v \quad \forall v \in X_h^1.$

Ainsi, si les espaces de discrétisation grossier et fin sont choisis de façon adéquate, l'erreur commise par l'approximation, à l'aide de ces schémas, sera du même ordre que celle commise lors de la résolution directe du problème non linéaire sur la grille fine. En effet, si la forme bilinéaire a et la fonction G sont telles que les problèmes 2 et 3 soient bien posés, il existe alors une constante C positive et δ_0 tels que, pour $0 < h < H \leq \delta_0$, on ait, pour chacun des problèmes présentés plus haut, une estimation de la forme suivante

$$\|u - v_h^H\|_{H^1(\Omega)} \quad \leq \quad C[h + H^2] \; \|u\|_{H^2(\Omega)},$$

où v_h^H désigne la solution du problème 1, 2 ou 3. Cette estimation est semblable à $\|u - u_h\|_{H^1(\Omega)} \leq Ch\|u\|_{H^2(\Omega)}$ dès lors que $h \sim H^2$.

Remarque : *Cette analyse n'a été faite que pour des grilles d'éléments finis* $\mathbb{P}_1$*, mais il est raisonnable de penser que cette méthode peut facilement s'adapter aux discrétisations de type spectrale, et plus particulièrement, à celle par les ondes planes. En effet, pour obtenir ces dernières estimations d'erreurs, il a fallu utiliser les résultats sur la convergence des approximations en norme* L^2*, et en norme négative, obtenus un peu plus tôt.*

Par ailleurs, ces schémas à deux grilles peuvent être également adaptés à la méthode de bases réduites.

Une méthode de bases réduites alternative

La méthode de bases réduites [34] est apparue dans les années 70, pour l'étude de problèmes de mécanique des solides non linéaires [40], et a été étendue à un grand nombre d'équations aux dérivées partielles dépendant de paramètres [36, 42]. Cette méthode repose sur le fait que l'ensemble des solutions $u(\mu)$, variant selon la valeur de μ, est de faible dimension. En effet, il existe un ensemble (de taille raisonnable) de paramètres μ_i, qui, si ils sont bien choisis, permettent d'approcher n'importe quelle solution $u(\mu)$ à l'aide d'une combinaison linéaire des solutions dépendant de ces μ_i, à un seuil de tolérance près. Cette méthode d'approximation est une alternative aux méthodes usuelles, car le nombre de degrés de liberté nécessaire est moins grand. En fait plus qu'une alternative c'est un complément car les éléments de la bases sont calculés par la méthode méthode des éléments finis. Il existe deux manières pour déterminer ses solutions particulières $u(\mu_i)$:

- les méthodes de décomposition propre orthogonale (POD)
- les algorithmes gloutons, bien que moins coûteux, nécessitant la définition d'indicateurs d'erreur *a posteriori* [33, 46].

L'implémentation de la méthode de bases réduites, se découpe en deux étapes :

- la première, dite « hors ligne », durant laquelle la base sera construite, et les matrices pourront être assemblées à l'aide de techniques de décomposition affine, ou d'interpolation sur « points magiques ». C'est la phase la plus coûteuse puisqu'on utilisera une méthode de discretisation habituelle (éléments finis, ou autres),
- la seconde étape, dite « en ligne », est plus rapide puisqu'elle ne fait intervenir que la base réduite.

Pour assembler rapidement la matrice de rigidité, pour chaque nouvelle valeur de μ, il existe deux méthodes :

- la décomposition affine des paramètres, si $a(u, v, \mu)$ peut s'écrire sous la forme $\sum_p \gamma(p) a_p(u, v)$,

– l'interpolation sur des « points magiques »pour traiter les non linéarités [18, 35].

Ainsi, il suffit de pré-calculer les parties indépendantes des paramètres durant l'étape « hors-ligne », pour gagner en temps de calcul. Pour mettre en oeuvre ces techniques, il faut avoir accès au code simulation. Comment faire alors, dans le cas industriel où les codes sont utilisés en boite noire ?
Il existe une alternative, exposée dans la partie 2, qui utilise des arguments de sous-grille d'éléments finis (inspiré de [20, 54]) et dont on peut accélérer la convergence grâce à des arguments de bases réduites et un post-traitement.

Plan de la thèse

Partie 1 :

Le but de cette première partie est de montrer que l'utilisation de schémas à deux grilles pour la résolution de problèmes aux valeurs propres non linéaires est possible. Pour cela, il faudra préalablement montrer que l'erreur mesurée en norme L^2 a un ordre plus élevé que celle mesurée en norme H^1. De nombreux travaux ont été effectués dans les cas de problèmes aux valeurs propres linéaires [2, 45, 50] mais il existe peu de résultats sur l'analyse numérique de problèmes aux valeurs propres non linéaires [28, 59, 60].

Dans le premier chapitre, des résultats de convergence seront établis pour le cas d'un problème simple. Une attention particulière sera portée aux normes négatives. En effet, l'utilisation de celles-ci permettra d'obtenir le même type d'estimation que dans le cas d'un problème aux valeurs propres linéaire. L'effet de l'intégration numérique devra également être pris en compte pour obtenir des taux de convergence optimaux. Cette étude sera faite sur deux types de discrétisations : spectrales (Fourier) et élements finis.

Le second chapitre est consacré à l'analyse numérique du modèle de Thomas-Fermi-Von-Weizacker.

Le dernier chapitre de cette partie sera dédié à l'analyse numérique de nouveaux schémas à deux grilles, pour la résolution de problèmes aux valeurs propres non linéaires.

Partie 2 :

La méthode des bases réduites permet la diminution du nombre de degrés de liberté dans l'approximation d'un système d'équations aux dérivées partielles dépendant d'un paramètre.
La procédure numérique de la méthode des bases réduites se déroule en deux étapes. Dans un premier temps, une étape de pré-calcul a lieu, dans laquelle

la base réduite et des fonctions associées sont calculées pour un ensemble de points prescrits dans l'espace des paramètres. Dans un deuxième temps, la solution approchée du modèle en temps réel est calculée pour une valeur quelconque du paramètre.
La première étape de cette procédure, nécessite l'accès au code de simulation, ce qui peut poser problème dans le cas de codes industriels. Ces derniers sont souvent utilisés en boite noire.
Le but de cette seconde partie est d'adapter les schémas à deux grilles aux méthodes de bases réduites. Ainsi, il sera possible d'obtenir des taux de convergence optimaux, tout en diminuant le nombre de degrés de liberté, et ce, même dans le cas d'un code utilisable uniquement en boite noire.
En effet, en combinant les propriétés de convergence des normes négatives et celles des bases réduites, ceci est possible. Le premier chapitre sera essentiellement consacré à l'introduction au méthodes de bases réduites, le second sera dédié à l'étude de cette nouvelle méthode.

Première partie

Schéma à deux grilles pour la résolution de problèmes aux valeurs propres non linéaires

Chapitre 1

Analyse numérique de problèmes aux valeurs propres non linéaires : un premier modèle

Numerical analysis of nonlinear eigenvalue problems

Eric Cancès[1], Rachida Chakir[2] and Yvon Maday[3]

[1] *Université Paris-Est, CERMICS, Project-team Micmac, INRIA-Ecole des Ponts, 6&8 avenue Blaise Pascal, 77455 Marne-la-Vallée Cedex 2, France.*

[2,3] *UPMC Univ Paris 06, UMR 7598 LJLL, Paris, F-75005 France ; CNRS, UMR 7598 LJLL, Paris, F-75005 France.*

[3] *Division of Applied Mathematics, Brown University, Providence, RI, USA.*

Abstract

We provide *a priori* error estimates for variational approximations of the ground state energy, eigenvalue and eigenvector of nonlinear elliptic eigenvalue problems of the form $-\mathrm{div}(A\nabla u) + Vu + f(u^2)u = \lambda u$, $\|u\|_{L^2} = 1$. We focus in particular on the Fourier spectral approximation (for periodic problems) and on the $\mathbb{P}_1$ and $\mathbb{P}_2$ finite-element discretizations. Denoting by $(u_\delta, \lambda_\delta)$ a variational approximation of the ground state eigenpair (u, λ), we are interested in the convergence rates of $\|u_\delta - u\|_{H^1}$, $\|u_\delta - u\|_{L^2}$, $|\lambda_\delta - \lambda|$, and the ground state energy, when the discretization parameter δ goes to zero. We prove in particular that if A, V and f satisfy certain conditions, $|\lambda_\delta - \lambda|$ goes to zero as $\|u_\delta - u\|_{H^1}^2 + \|u_\delta - u\|_{L^2}$. We also show that under more restrictive assumptions on A, V and f, $|\lambda_\delta - \lambda|$ converges to zero as $\|u_\delta - u\|_{H^1}^2$, thus recovering a standard result for *linear* elliptic eigenvalue problems. For the latter analysis, we make use of estimates of the error $u_\delta - u$ in negative Sobolev norms.

Keywords: nonlinear eigenvalue problem, convergence analysis, superconvergence, Fourier spectral approximation, finite element approximation.

1 Introduction

Many mathematical models in science and engineering give rise to nonlinear eigenvalue problems. Let us mention for instance the calculation of the vibration modes of a mechanical structure in the framework of nonlinear elasticity, the Gross-Pitaevskii equation describing the steady states of Bose-Einstein condensates [38], or the Hartree-Fock and Kohn-Sham equations used to calculate ground state electronic structures of molecular systems in quantum chemistry and materials science (see [9] for a mathematical introduction).

The numerical analysis of *linear* eigenvalue problems has been thoroughly studied in the past decades (see e.g. [2]). On the other hand, only a few results on *nonlinear* eigenvalue problems have been published so far [59, 60].

In this article, we focus on a particular class of nonlinear eigenvalue problems arising in the study of variational models of the form

$$I = \inf \left\{ E(v),\ v \in X,\ \int_\Omega v^2 = 1 \right\} \tag{1}$$

where

Ω is a regular bounded domain or a rectangular brick of $\mathbb{R}^d$ and $X = H^1_0(\Omega)$
or
Ω is the unit cell of a periodic lattice $\mathcal{R}$ of $\mathbb{R}^d$ and $X = H^1_\#(\Omega)$

with $d = 1$, 2 or 3, and where the energy functional E is of the form

$$E(v) = \frac{1}{2} a(v,v) + \frac{1}{2} \int_\Omega F(v^2(x))\, dx$$

with

$$a(u,v) = \int_\Omega (A\nabla u) \cdot \nabla v + \int_\Omega V uv.$$

Recall that if Ω is the unit cell of a periodic lattice $\mathcal{R}$ of $\mathbb{R}^d$, then for all $s \in \mathbb{R}$ and $k \in \mathbb{N}$,

$$\begin{aligned} H^s_\#(\Omega) &= \left\{ v|_\Omega,\ v \in H^s_{\rm loc}(\mathbb{R}^d) \mid v\ \mathcal{R}\text{-periodic} \right\}, \\ C^k_\#(\Omega) &= \left\{ v|_\Omega,\ v \in C^k(\mathbb{R}^d) \mid v\ \mathcal{R}\text{-periodic} \right\}. \end{aligned}$$

We assume in addition that

- $A \in (L^\infty(\Omega))^{d\times d}$ and $A(x)$ is symmetric for almost all $x \in \Omega$ (2)
  $\exists \alpha > 0$ s.t. $\xi^T A(x) \xi \geq \alpha |\xi|^2$ for all $\xi \in \mathbb{R}^d$ and almost all $x \in \Omega$ (3)

- $V \in L^p(\Omega)$ for some $p > \max(1, d/2)$ (4)

- $F \in C^1([0,+\infty), \mathbb{R}) \cap C^2((0,\infty), \mathbb{R})$ and $F'' > 0$ on $(0,+\infty)$ (5)
  $\exists 0 \leq q < 2,\ \exists C \in \mathbb{R}_+$ s.t. $\forall t \geq 0,\ |F'(t)| \leq C(1 + t^q)$ (6)
  $F''(t)t$ remains bounded in the vicinity of 0. (7)

To establish some of our results, we will also need to make the additional assumption that there exists $1 < r \leq 2$ and $0 \leq s \leq 5 - r$ such that

$$\begin{aligned} &\forall R > 0,\ \exists C_R \in \mathbb{R}_+ \text{ s.t. } \forall 0 < t_1 \leq R,\ \forall t_2 \in \mathbb{R}, \\ &\quad \left| F'(t_2^2)t_2 - F'(t_1^2)t_2 - 2F''(t_1^2)t_1^2(t_2 - t_1) \right| \leq C_R \left(1 + |t_2|^s\right) |t_2 - t_1|^r. \end{aligned} \tag{8}$$

Note that for all $1 < m < 3$ and all $c > 0$, the function $F(t) = ct^m$ satisfies (5)-(7) and (8), for some $1 < r \leq 2$. It satisfies (8) with $r = 2$ if $3/2 \leq m < 3$. This allows us to handle the Thomas-Fermi kinetic energy functional ($m = \frac{5}{3}$) as well as the repulsive interaction in Bose-Einstein condensates ($m = 2$).

Remark 1.1 *Assumption (6) is sharp for $d = 3$, but is useless for $d = 1$ and can be replaced with the weaker assumption that there exist $q < \infty$ and $C \in \mathbb{R}_+$ such that $|F'(t)| \leq C(1 + t^q)$ for all $t \in \mathbb{R}_+$, for $d = 2$. Likewise, the condition $0 \leq s \leq 5 - r$ in assumption (8) is sharp for $d = 3$ but can be replaced with $0 \leq s < \infty$ if $d = 1$ or $d = 2$.*

In order to simplify the notation, we denote by $f(t) = F'(t)$.

Making the change of variable $\rho = v^2$ and noticing that $a(|v|, |v|) = a(v, v)$ for all $v \in X$, it is easy to check that

$$I = \inf \left\{ \mathcal{E}(\rho),\ \rho \geq 0,\ \sqrt{\rho} \in X,\ \int_\Omega \rho = 1 \right\}, \tag{9}$$

where

$$\mathcal{E}(\rho) = \frac{1}{2} a(\sqrt{\rho}, \sqrt{\rho}) + \frac{1}{2} \int_\Omega F(\rho).$$

We will see that under assumptions (2)-(6), (9) has a unique solution ρ_0 and (1) has exactly two solutions: $u = \sqrt{\rho_0}$ and $-u$. Moreover, E is C^1 on X and for all $v \in X$, $E'(v) = A_v v$ where

$$A_v = -\mathrm{div}(A\nabla\cdot) + V + f(v^2).$$

Note that A_v defines a self-adjoint operator on $L^2(\Omega)$, with form domain X. The function u therefore is solution to the Euler equation

$$\forall v \in X, \quad \langle A_u u - \lambda u, v \rangle_{X',X} = 0 \tag{10}$$

for some $\lambda \in \mathbb{R}$ (the Lagrange multiplier of the constraint $\|u\|_{L^2}^2 = 1$) and equation (10), complemented with the constraint $\|u\|_{L^2} = 1$, takes the form of the nonlinear eigenvalue problem

$$\left\{ \begin{array}{l} A_u u = \lambda u \\ \|u\|_{L^2} = 1. \end{array} \right. \tag{11}$$

In addition, $u \in C^0(\overline{\Omega})$, $u > 0$ in Ω and λ is the ground state eigenvalue of the linear operator A_u. An important result is that λ is a *simple* eigenvalue of A_u. It is interesting to note that λ is also the ground state eigenvalue of the *nonlinear* eigenvalue problem

$$\left\{ \begin{array}{l} \text{search } (\mu, v) \in \mathbb{R} \times X \text{ such that} \\ A_v v = \mu v \\ \|v\|_{L^2} = 1, \end{array} \right. \tag{12}$$

in the following sense: if (μ, v) is solution to (12) then either $\mu > \lambda$ or $\mu = \lambda$ and $v = \pm u$. All these properties, except maybe the last one, are classical. For the sake of completeness, their proofs are however given in the Appendix.

Let us now turn to the main topic of this article, namely the derivation of a priori error estimates for variational approximations of the ground state

eigenpair (λ, u). We denote by $(X_\delta)_{\delta>0}$ a family of finite-dimensional subspaces of X such that

$$\min\{\|u - v_\delta\|_{H^1},\ v_\delta \in X_\delta\} \underset{\delta\to 0^+}{\longrightarrow} 0 \tag{13}$$

and consider the variational approximation of (1) consisting in solving

$$I_\delta = \inf\left\{E(v_\delta),\ v_\delta \in X_\delta,\ \int_\Omega v_\delta^2 = 1\right\}. \tag{14}$$

Problem (14) has at least one minimizer u_δ, which satisfies

$$\forall v_\delta \in X_\delta, \quad \langle A_{u_\delta} u_\delta - \lambda_\delta u_\delta, v_\delta\rangle_{X',X} = 0 \tag{15}$$

for some $\lambda_\delta \in \mathbb{R}$. Obviously, $-u_\delta$ also is a minimizer associated with the same eigenvalue λ_δ. On the other hand, it is not known whether u_δ and $-u_\delta$ are the only minimizers of (14). One of the reasons why the argument used in the infinite-dimensional setting cannot be transposed to the discrete case is that the set

$$\left\{\rho \mid \exists u_\delta \in X_\delta \text{ s.t. } \|u_\delta\|_{L^2} = 1,\ \rho = u_\delta^2\right\}$$

is not convex in general. We will see however (cf. Theorem 1) that for any family $(u_\delta)_{\delta>0}$ of global minimizers of (14) such that $(u, u_\delta) \geq 0$ for all $\delta > 0$, the following holds true

$$\|u_\delta - u\|_{H^1} \underset{\delta\to 0^+}{\longrightarrow} 0.$$

In addition, a simple calculation leads to

$$\lambda_\delta - \lambda = \langle (A_u - \lambda)(u_\delta - u), (u_\delta - u)\rangle_{X',X} + \int_\Omega w_{u,u_\delta}(u_\delta - u) \tag{16}$$

where

$$w_{u,u_\delta} = u_\delta^2 \frac{f(u_\delta^2) - f(u^2)}{u_\delta - u}.$$

The first term of the right-hand side of (16) is nonnegative and goes to zero as $\|u_\delta - u\|_{H^1}^2$. We will prove in Theorem 1 that the second term goes to zero at least as $\|u_\delta - u\|_{L^{6/(5-2q)}}$. Therefore, $|\lambda_\delta - \lambda|$ converges to zero with δ at least as $\|u_\delta - u\|_{H^1}^2 + \|u_\delta - u\|_{L^{6/(5-2q)}}$.

The purpose of this article is to provide more precise *a priori* error bounds on $|\lambda_\delta - \lambda|$, as well as on $\|u_\delta - u\|_{H^1}$, $\|u_\delta - u\|_{L^2}$ and $E(u_\delta) - E(u)$. In Section 2, we prove a series of estimates valid in the general framework described above. We then turn to more specific examples, where the analysis can be pushed further. In Section 2, we concentrate on the discretization of problem (1) with

$$\begin{aligned}
&\Omega = (0, 2\pi)^d,\\
&X = H^1_{\#}(0, 2\pi)^d,\\
&E(v) = \frac{1}{2}\int_\Omega |\nabla v|^2 + \frac{1}{2}\int_\Omega V v^2 + \frac{1}{2}\int_\Omega F(v^2),
\end{aligned}$$

in Fourier modes. In Section 4, we deal with the $\mathbb{P}_1$ and $\mathbb{P}_2$ finite element discretizations of problem (1) with

$$\begin{aligned}&\Omega \text{ rectangular brick of } \mathbb{R}^d,\\ &X = H_0^1(\Omega),\\ &E(v) = \frac{1}{2}\int_\Omega |\nabla v|^2 + \frac{1}{2}\int_\Omega V v^2 + \frac{1}{2}\int_\Omega F(v^2).\end{aligned}$$

Lastly, we discuss the issue of numerical integration in Section 5.

2 Basic error analysis

The aim of this section is to establish error bounds on $\|u_\delta - u\|_{H^1}$, $\|u_\delta - u\|_{L^2}$, $|\lambda_\delta - \lambda|$ and $E(u_\delta) - E(u)$, in a general framework. In the whole section, we make the assumptions (2)-(7) and (13), and we denote by u the unique positive solution of (1) and by u_δ a minimizer of the discretized problem (14) such that $(u_\delta, u)_{L^2} \geq 0$. We also introduce the bilinear form $E''(u)$ defined on $X \times X$ by

$$\langle E''(u)v, w\rangle_{X',X} = \langle A_u v, w\rangle_{X',X} + 2\int_\Omega f'(u^2)u^2 vw.$$

When $F \in C^2([0,+\infty), \mathbb{R})$, then E is twice differentiable on X and $E''(u)$ is the second derivative of E at u.

Lemma 1 *There exists $\beta > 0$ and $M \in \mathbb{R}_+$ such that for all $v \in X$,*

$$0 \leq \langle (A_u - \lambda)v, v\rangle_{X',X} \leq M\|v\|_{H^1}^2 \tag{17}$$

$$\beta\|v\|_{H^1}^2 \leq \langle (E''(u) - \lambda)v, v\rangle_{X',X} \leq M\|v\|_{H^1}^2. \tag{18}$$

There exists $\gamma > 0$ such that for all $\delta > 0$,

$$\gamma\|u_\delta - u\|_{H^1}^2 \leq \langle (A_u - \lambda)(u_\delta - u), (u_\delta - u)\rangle_{X',X}. \tag{19}$$

Proof We have for all $v \in X$,

$$\langle (A_u - \lambda)v, v\rangle_{X',X} \leq \|A\|_{L^\infty}\|\nabla v\|_{L^2}^2 + \|V\|_{L^p}\|v\|_{L^{2p'}}^2 + \|f(u^2)\|_{L^\infty}\|v\|_{L^2}^2$$

where $p' = (1 - p^{-1})^{-1}$ and

$$\langle (E''(u) - \lambda)v, v\rangle_{X',X} \leq \langle (A_u - \lambda)v, v\rangle_{X',X} + 2\|f'(u^2)u^2\|_{L^\infty}\|v\|_{L^2}^2.$$

Hence the upper bounds in (17) and (18). We now use the fact that λ, the lowest eigenvalue of A_u, is simple (see Lemma 2 in the Appendix). This implies that there exists $\eta > 0$ such that

$$\forall v \in X, \quad \langle (A_u - \lambda)v, v\rangle_{X',X} \geq \eta(\|v\|_{L^2}^2 - |(u,v)_{L^2}|^2) \geq 0. \tag{20}$$

This provides on the one hand the lower bound (17), and leads on the other hand to the inequality

$$\forall v \in X, \quad \langle (E''(u) - \lambda)v, v \rangle_{X',X} \geq 2 \int_\Omega f'(u^2) u^2 v^2.$$

As $f' = F'' > 0$ in $(0, +\infty)$ and $u > 0$ in Ω, we therefore have

$$\forall v \in X \setminus \{0\}, \quad \langle (E''(u) - \lambda)v, v \rangle_{X',X} > 0.$$

Reasoning by contradiction, we deduce from the above inequality and the first inequality in (20) that there exists $\widetilde{\eta} > 0$ such that

$$\forall v \in X, \quad \langle (E''(u) - \lambda)v, v \rangle_{X',X} \geq \widetilde{\eta} \|v\|_{L^2}^2. \tag{21}$$

Besides, there exists a constant $C \in \mathbb{R}_+$ such that

$$\forall v \in X, \quad \langle (A_u - \lambda)v, v \rangle_{X',X} \geq \frac{\alpha}{2} \|\nabla v\|_{L^2}^2 - C\|v\|_{L^2}^2. \tag{22}$$

Let us establish this inequality for $d = 3$ (the case when $d = 1$ is straightforward and the case when $d = 2$ can be dealt with in the same way). For all $x \in X$,

$$\begin{aligned}
\langle (A_u - \lambda)v, v \rangle_{X',X} &= \int_\Omega (A\nabla v) \cdot \nabla v + \int_\Omega (V + f(v^2) - \lambda) v^2 \\
&\geq \alpha \|\nabla v\|_{L^2}^2 - \|V\|_{L^p} \|v\|_{L^{2p'}}^2 + (f(0) - \lambda)\|v\|_{L^2}^2 \\
&\geq \alpha \|\nabla v\|_{L^2}^2 - \|V\|_{L^p} \|v\|_{L^2}^{2-3/p} \|v\|_{L^6}^{3/p} + (f(0) - \lambda)\|v\|_{L^2}^2 \\
&\geq \alpha \|\nabla v\|_{L^2}^2 - C_6^{3/p} \|V\|_{L^p} \|v\|_{L^2}^{2-3/p} \|v\|_{H^1}^{3/p} + (f(0) - \lambda)\|v\|_{L^2}^2 \\
&\geq \frac{\alpha}{2} \|\nabla v\|_{L^2}^2 \\
&\quad + \left(f(0) - \lambda - \frac{3 - 2p}{2p} \left(\frac{3C_6^2 \|V\|_{L^p}^{2p/3}}{p\alpha} \right)^{3/(2p-3)} - \frac{\alpha}{2} \right) \|v\|_{L^2}^2,
\end{aligned}$$

where C_6 is the Sobolev constant such that $\forall v \in X$, $\|v\|_{L^6} \leq C_6 \|v\|_{H^1}$. The coercivity of $E''(u) - \lambda$ (i.e. the lower bound in (18)) is a straightforward consequence of (21) and (22).

To prove (19), we notice that

$$\|u_\delta\|_{L^2}^2 - |(u, u_\delta)_{L^2}|^2 \geq 1 - (u, u_\delta)_{L^2} = \frac{1}{2} \|u_\delta - u\|_{L^2}^2.$$

It therefore readily follows from (20) that

$$\langle (A_u - \lambda)(u_\delta - u), (u_\delta - u) \rangle_{X',X} \geq \frac{\eta}{2} \|u_\delta - u\|_{L^2}^2.$$

Combining with (22), we finally obtain (19). □

For $w \in X'$, we denote by ψ_w the unique solution to the adjoint problem

$$\left\{ \begin{array}{l} \text{find } \psi_w \in u^\perp \text{ such that} \\ \forall v \in u^\perp, \quad \langle (E''(u) - \lambda)\psi_w, v\rangle_{X',X} = \langle w, v\rangle_{X',X}, \end{array} \right. \tag{23}$$

where

$$u^\perp = \left\{ v \in X \mid \int_\Omega uv = 0 \right\}.$$

The existence and uniqueness of the solution to (64) is a straightforward consequence of (18) and the Lax-Milgram lemma. Besides,

$$\forall w \in L^2(\Omega), \quad \|\psi_w\|_{H^1} \leq \beta^{-1} M \|w\|_{X'} \leq \beta^{-1} M \|w\|_{L^2}. \tag{24}$$

We can now state the main result of this section.

Theorem 1 *Under assumptions (2)-(6) and (13), it holds*

$$\|u_\delta - u\|_{H^1} \underset{\delta \to 0^+}{\longrightarrow} 0.$$

If in addition, (7) is satisfied, then there exists $C \in \mathbb{R}_+$ such that for all $\delta > 0$,

$$\frac{\gamma}{2}\|u_\delta - u\|_{H^1}^2 \leq E(u_\delta) - E(u) \leq \frac{M}{2}\|u_\delta - u\|_{H^1}^2 + C\|u_\delta - u\|_{L^{6/(5-2q)}}, \tag{25}$$

and

$$|\lambda_\delta - \lambda| \leq C\left(\|u_\delta - u\|_{H^1}^2 + \|u_\delta - u\|_{L^{6/(5-2q)}}\right). \tag{26}$$

Besides, if assumption (8) is satisfied for some $1 < r \leq 2$ and $0 \leq s \leq 5 - r$, then there exists $\delta_0 > 0$ and $C \in \mathbb{R}_+$ such that for all $0 < \delta < \delta_0$,

$$\|u_\delta - u\|_{H^1} \leq C \min_{v_\delta \in X_\delta} \|v_\delta - u\|_{H^1} \tag{27}$$

$$\|u_\delta - u\|_{L^2}^2 \leq C\left(\|u_\delta - u\|_{L^2}\|u_\delta - u\|_{L^{6r/(5-s)}}^r + \|u_\delta - u\|_{H^1} \min_{\psi_\delta \in X_\delta} \|\psi_{u_\delta - u} - \psi_\delta\|_{H^1} \right). \tag{28}$$

Lastly, if F'' is bounded in the vicinity of 0, there exists $C \in \mathbb{R}_+$ such that for all $\delta > 0$,

$$\frac{\gamma}{2}\|u_\delta - u\|_{H^1}^2 \leq E(u_\delta) - E(u) \leq C\|u_\delta - u\|_{H^1}^2. \tag{29}$$

Remark 2.1 *If $0 \leq r + s \leq 3$, then*

$$\|u_\delta - u\|_{L^{6r/(5-s)}}^r \leq \|u_\delta - u\|_{L^2}^{(5-r-s)/2}\|u_\delta - u\|_{L^6}^{(3r-5+s)/2} \leq \|u_\delta - u\|_{L^2}\|u_\delta - u\|_{H^1}^{r-1},$$

so that (70) implies the simpler inequality

$$\|u_\delta - u\|_{L^2}^2 \leq C\|u_\delta - u\|_{H^1} \min_{\psi_\delta \in X_\delta} \|\psi_{u_\delta - u} - \psi_\delta\|_{H^1}. \tag{30}$$

Proof of Theorem 1 We have

$$\begin{aligned} E(u_\delta) - E(u) &= \frac{1}{2}\langle A_u u_\delta, u_\delta\rangle_{X',X} - \frac{1}{2}\langle A_u u, u\rangle_{X',X} \\ &\quad + \frac{1}{2}\int_\Omega F\left(u_\delta^2\right) - F\left(u^2\right) - f\left(u^2\right)\left(u_\delta^2 - u^2\right) \\ &= \frac{1}{2}\langle (A_u - \lambda)(u_\delta - u), (u_\delta - u)\rangle_{X',X} \\ &\quad + \frac{1}{2}\int_\Omega F\left(u_\delta\right) - F\left(u^2\right) - f\left(u^2\right)\left(u_\delta^2 - u^2\right). \end{aligned} \tag{31}$$

Using (19) and the convexity of F, we get

$$E(u_\delta) - E(u) \geq \frac{\gamma}{2}\|u_\delta - u\|_{H^1}^2.$$

Let $\Pi_\delta u \in X_\delta$ be such that

$$\|u - \Pi_\delta u\|_{H^1} = \min\left\{\|u - v_\delta\|_{H^1}, v_\delta \in X_\delta\right\}.$$

We deduce from (13) that $(\Pi_\delta u)_{\delta>0}$ converges to u in X when δ goes to zero. Denoting by $\widetilde{u}_\delta = \|\Pi_\delta u\|_{L^2}^{-1}\Pi_\delta u$ (which is well defined, at least for δ small enough), we also have

$$\lim_{\delta\to 0^+}\|\widetilde{u}_\delta - u\|_{H^1} = 0.$$

The functional E being strongly continuous on X, we obtain

$$\|u_\delta - u\|_{H^1}^2 \leq \frac{2}{\gamma}\left(E(u_\delta) - E(u)\right) \leq \frac{2}{\gamma}\left(E(\widetilde{u}_\delta) - E(u)\right) \underset{\delta\to 0^+}{\longrightarrow} 0.$$

It follows that there exists $\delta_1 > 0$ such that

$$\forall 0 < \delta \leq \delta_1, \quad \|u_\delta\|_{H^1} \leq 2\|u\|_{H^1}, \quad \|u_\delta - u\|_{H^1} \leq \frac{1}{2}.$$

We then easily deduce from (31) the upper bounds in (67) and (29).

Next, we remark that

$$\begin{aligned} \lambda_\delta - \lambda &= \langle E'(u_\delta), u_\delta\rangle_{X',X} - \langle E'(u), u\rangle_{X',X} \\ &= a(u_\delta, u_\delta) - a(u,u) + \int_\Omega f(u_\delta^2)u_\delta^2 - \int_\Omega f(u^2)u^2 \\ &= a(u_\delta - u, u_\delta - u) + 2a(u, u_\delta - u) + \int_\Omega f(u_\delta^2)u_\delta^2 - \int_\Omega f(u^2)u^2 \\ &= a(u_\delta - u, u_\delta - u) + 2\lambda\int_\Omega u(u_\delta - u) - 2\int_\Omega f(u^2)u(u_\delta - u) \\ &\qquad + \int_\Omega f(u_\delta^2)u_\delta^2 - \int_\Omega f(u^2)u^2 \\ &= a(u_\delta - u, u_\delta - u) - \lambda\|u_\delta - u\|_{L^2}^2 - 2\int_\Omega f(u^2)u(u_\delta - u) \\ &\qquad + \int_\Omega f(u_\delta^2)u_\delta^2 - \int_\Omega f(u^2)u^2 \\ &= \langle (A_u - \lambda)(u_\delta - u), (u_\delta - u)\rangle_{X',X} + \int_\Omega w_{u,u_\delta}(u_\delta - u) \end{aligned} \tag{32}$$

where

$$w_{u,u_\delta} = u_\delta^2 \frac{f(u_\delta^2) - f(u^2)}{u_\delta - u}.$$

As $u \in L^\infty(\Omega)$, we have

$$|w_{u,u_\delta}| \leq \begin{cases} 12u \sup_{t\in(0,4\|u\|^2_{L^\infty}]} F''(t)t & \text{if } |u_\delta| < 2u \\ 2\left(|f(u_\delta^2)| + \max_{t\in[0,\|u\|^2_{L^\infty}]} |f(t)|\right)|u_\delta| & \text{if } |u_\delta| \geq 2u, \end{cases}$$

and we deduce from assumptions (6)-(7) that

$$|w_{u,u_\delta}| \leq C(1 + |u_\delta|^{2q+1}),$$

for some constant C independent of δ. Using (17), we therefore obtain that for all $0 < \delta \leq \delta_1$,

$$\begin{aligned} |\lambda_\delta - \lambda| &\leq M\|u_\delta - u\|^2_{H^1} + \|w_{u,u_\delta}\|_{L^{6/(2q+1)}}\|u_\delta - u\|_{L^{6/(5-2q)}} \\ &\leq M\|u_\delta - u\|^2_{H^1} + C(1 + \|u_\delta\|^{2q+1}_{H^1})\|u_\delta - u\|_{L^{6/(5-2q)}} \\ &\leq C\left(\|u_\delta - u\|^2_{H^1} + \|u_\delta - u\|_{L^{6/(5-2q)}}\right), \end{aligned} \tag{33}$$

where C denotes constants independent of δ.

In order to evaluate the H^1-norm of the error $u_\delta - u$, we first notice that

$$\forall v_\delta \in X_\delta, \quad \|u_\delta - u\|_{H^1} \leq \|u_\delta - v_\delta\|_{H^1} + \|v_\delta - u\|_{H^1}, \tag{34}$$

and that

$$\begin{aligned} \|u_\delta - v_\delta\|^2_{H^1} &\leq \beta^{-1}\langle (E''(u) - \lambda)(u_\delta - v_\delta), (u_\delta - v_\delta)\rangle_{X',X} \\ &= \beta^{-1}\Big(\langle (E''(u) - \lambda)(u_\delta - u), (u_\delta - v_\delta)\rangle_{X',X} \\ &\qquad + \langle (E''(u) - \lambda)(u - v_\delta), (u_\delta - v_\delta)\rangle_{X',X}\Big). \end{aligned} \tag{35}$$

For all $w_\delta \in X_\delta$

$$\begin{aligned} &\langle (E''(u) - \lambda)(u_\delta - u), w_\delta\rangle_{X',X} \\ &= -\int_\Omega \left(f(u_\delta^2)u_\delta - f(u^2)u_\delta - 2f'(u^2)u^2(u_\delta - u)\right) w_\delta + (\lambda_\delta - \lambda)\int_\Omega u_\delta w_\delta. \end{aligned} \tag{36}$$

On the other hand, we have for all $v_\delta \in X_\delta$ such that $\|v_\delta\|_{L^2} = 1$,

$$\int_\Omega u_\delta(u_\delta - v_\delta) = 1 - \int_\Omega u_\delta v_\delta = \frac{1}{2}\|u_\delta - v_\delta\|^2_{L^2}.$$

Using (8) and (33), we therefore obtain that for all $0 < \delta \leq \delta_1$ and all $v_\delta \in X_\delta$ such that $\|v_\delta\|_{L^2} = 1$,

$$\begin{aligned} \left|\langle (E''(u) - \lambda)(u_\delta - u), (u_\delta - v_\delta)\rangle_{X',X}\right| &\leq C\Big(\|u_\delta - u\|^r_{L^{6r/(5-s)}}\|u_\delta - v_\delta\|_{H^1} \\ &\quad + \left(\|u_\delta - u\|^2_{H^1} + \|u_\delta - u\|_{L^{6/(5-2q)}}\right)\|u_\delta - v_\delta\|^2_{L^2}\Big). \end{aligned} \tag{37}$$

It then follows from (18), (35) and (78) that for all $0 < \delta \leq \delta_1$ and all $v_\delta \in X_\delta$ such that $\|v_\delta\|_{L^2} = 1$,

$$\|u_\delta - v_\delta\|_{H^1} \leq C\left(\|u_\delta - u\|_{H^1}^r + \|u_\delta - u\|_{H^1}\|u_\delta - v_\delta\|_{H^1} + \|v_\delta - u\|_{H^1}\right).$$

Combining with (73) we obtain that there exists $0 < \delta_2 \leq \delta_1$ and $C \in \mathbb{R}_+$ such that for all $0 < \delta \leq \delta_2$ and all $v_\delta \in X_\delta$ such that $\|v_\delta\|_{L^2} = 1$,

$$\|u_\delta - u\|_{H^1} \leq C\|v_\delta - u\|_{H^1}.$$

Hence, for all $0 < \delta \leq \delta_2$

$$\|u_\delta - u\|_{H^1} \leq CJ_\delta \qquad \text{where} \qquad J_\delta = \min_{v_\delta \in X_\delta \,|\, \|v_\delta\|_{L^2}=1} \|v_\delta - u\|_{H^1}.$$

We now denote by

$$\widetilde{J}_\delta = \min_{v_\delta \in X_\delta} \|v_\delta - u\|_{H^1},$$

and by u_δ^0 a minimizer of the above minimization problem. We know from (13) that u_δ^0 converges to u in H^1 when δ goes to zero. Besides,

$$\begin{aligned} J_\delta &\leq \|u_\delta^0/\|u_\delta^0\|_{L^2} - u\|_{H^1} \\ &\leq \|u_\delta^0 - u\|_{H^1} + \frac{\|u_\delta^0\|_{H^1}}{\|u_\delta^0\|_{L^2}}\left|1 - \|u_\delta^0\|_{L^2}\right| \\ &\leq \|u_\delta^0 - u\|_{H^1} + \frac{\|u_\delta^0\|_{H^1}}{\|u_\delta^0\|_{L^2}}\|u - u_\delta^0\|_{L^2} \\ &\leq \left(1 + \frac{\|u_\delta^0\|_{H^1}}{\|u_\delta^0\|_{L^2}}\right)\widetilde{J}_\delta. \end{aligned}$$

For $0 < \delta \leq \delta_2 \leq \delta_1$, we have $\|u_\delta^0 - u\|_{H^1} \leq \|u_\delta - u\|_{H^1} \leq 1/2$, and therefore $\|u_\delta^0\|_{H^1} \leq \|u\|_{H^1} + 1/2$ and $\|u_\delta^0\|_{L^2} \geq 1/2$, yielding $J_\delta \leq 2(\|u\|_{H^1} + 1)\widetilde{J}_\delta$. Thus (69) is proved.

Let u_δ^* be the orthogonal projection, for the L^2 inner product, of u_δ on the affine space $\left\{v \in L^2(\Omega) \,|\, \int_\Omega uv = 1\right\}$. One has

$$u_\delta^* \in X, \qquad u_\delta^* - u \in u^\perp, \qquad u_\delta^* - u_\delta = \frac{1}{2}\|u_\delta - u\|_{L^2}^2 u,$$

from which we infer that

$$\begin{aligned}
\|u_\delta - u\|_{L^2}^2 &= \int_\Omega (u_\delta - u)(u_\delta^* - u) + \int_\Omega (u_\delta - u)(u_\delta - u_\delta^*) \\
&= \int_\Omega (u_\delta - u)(u_\delta^* - u) - \frac{1}{2}\|u_\delta - u\|_{L^2}^2 \int_\Omega (u_\delta - u)u \\
&= \int_\Omega (u_\delta - u)(u_\delta^* - u) + \frac{1}{2}\|u_\delta - u\|_{L^2}^2 \left(1 - \int_\Omega u_\delta u\right) \\
&= \int_\Omega (u_\delta - u)(u_\delta^* - u) + \frac{1}{4}\|u_\delta - u\|_{L^2}^4 \\
&= \langle u_\delta - u, u_\delta^* - u\rangle_{X',X} + \frac{1}{4}\|u_\delta - u\|_{L^2}^4 \\
&= \langle (E''(u) - \lambda)\psi_{u_\delta - u}, u_\delta^* - u\rangle_{X',X} + \frac{1}{4}\|u_\delta - u\|_{L^2}^4 \\
&= \langle (E''(u) - \lambda)(u_\delta - u), \psi_{u_\delta - u}\rangle_{X',X} \\
&\quad + \frac{1}{2}\|u_\delta - u\|_{L^2}^2 \langle (E''(u) - \lambda)u, \psi_{u_\delta - u}\rangle_{X',X} + \frac{1}{4}\|u_\delta - u\|_{L^2}^4 \\
&= \langle (E''(u) - \lambda)(u_\delta - u), \psi_{u_\delta - u}\rangle_{X',X} \\
&\quad + \|u_\delta - u\|_{L^2}^2 \int_\Omega f'(u^2)u^3\psi_{u_\delta - u} + \frac{1}{4}\|u_\delta - u\|_{L^2}^4.
\end{aligned}$$

For all $\psi_\delta \in X_\delta$, it therefore holds

$$\begin{aligned}
\|u_\delta - u\|_{L^2}^2 &= \langle (E''(u) - \lambda)(u_\delta - u), \psi_\delta\rangle_{X',X} \\
&\quad + \langle (E''(u) - \lambda)(u_\delta - u), \psi_{u_\delta - u} - \psi_\delta\rangle_{X',X} \\
&\quad + \|u_\delta - u\|_{L^2}^2 \int_\Omega f'(u^2)u^3\psi_{u_\delta - u} + \frac{1}{4}\|u_\delta - u\|_{L^2}^4.
\end{aligned}$$

From (75), we obtain that for all $\psi_\delta \in X_\delta \cap u^\perp$,

$$\begin{aligned}
&\langle (E''(u) - \lambda)(u_\delta - u), \psi_\delta\rangle_{X',X} \\
&= -\int_\Omega \left(f(u_\delta^2)u_\delta - f(u^2)u_\delta - 2f'(u^2)u^2(u_\delta - u)\right)\psi_\delta + (\lambda_\delta - \lambda)\int_\Omega (u_\delta - u)\psi_\delta
\end{aligned}$$

and therefore that for all $\psi_\delta \in X_\delta \cap u^\perp$,

$$\begin{aligned}
\left|\langle (E''(u) - \lambda)(u_\delta - u), \psi_\delta\rangle_{X',X}\right| &\le C\Big(\|u_\delta - u\|_{L^{6r/(5-s)}}^r \\
&\quad + \|u_\delta - u\|_{L^{6/5}}\left(\|u_\delta - u\|_{H^1}^2 + \|u_\delta - u\|_{L^{6/(5-2q)}}\right)\Big)\|\psi_\delta\|_{H^1} \quad (38)
\end{aligned}$$

Let $\psi_\delta^0 \in X_\delta \cap u^\perp$ be such that

$$\|\psi_{u_\delta - u} - \psi_\delta^0\|_{H^1} = \min_{\psi_\delta \in X_\delta \cap u^\perp} \|\psi_{u_\delta - u} - \psi_\delta\|_{H^1}.$$

Noticing that $\|\psi_\delta^0\|_{H^1} \le \|\psi_{u_\delta - u}\|_{H^1} \le \beta^{-1}M\|u_\delta - u\|_{L^2}$, we obtain from (18)

and (82) that there exists $C \in \mathbb{R}_+$ such that for all $0 < \delta \leq \delta_1$,

$$\begin{aligned}\|u_\delta - u\|_{L^2}^2 &\leq C\Big(\|u_\delta - u\|_{L^2} \\ &\quad \times \left(\|u_\delta - u\|_{L^{6r/(5-s)}}^r + \|u_\delta - u\|_{L^{6/5}} \left(\|u_\delta - u\|_{H^1}^2 + \|u_\delta - u\|_{L^{6/(5-2q)}} \right) \right) \\ &\quad + \|u_\delta - u\|_{H^1} \|\psi_{u_\delta - u} - \psi_\delta^0\|_{H^1} + \|u_\delta - u\|_{L^2}^3 + \|u_\delta - u\|_{L^2}^4 \Big).\end{aligned}$$

Therefore, there exists $0 < \delta_0 \leq \delta_2$ and $C \in \mathbb{R}_+$ such that for all $0 < \delta \leq \delta_0$,

$$\|u_\delta - u\|_{L^2}^2 \leq C \left(\|u_\delta - u\|_{L^2} \|u_\delta - u\|_{L^{6r/(5-s)}}^r + \|u_\delta - u\|_{H^1} \|\psi_{u_\delta - u} - \psi_\delta^0\|_{H^1} \right).$$

Lastly, denoting by $\Pi^0_{X_\delta}$ the orthogonal projector on X_δ for the L^2 inner product, a simple calculation leads to

$$\forall v \in u^\perp, \quad \min_{v_\delta \in X_\delta \cap u^\perp} \|v_\delta - v\|_{H^1} \leq \left(1 + \frac{\|\Pi^0_{X_\delta} u\|_{H^1}}{\|\Pi^0_{X_\delta} u\|_{L^2}^2} \right) \min_{v_\delta \in X_\delta} \|v_\delta - v\|_{H^1}, \quad (39)$$

which completes the proof of Theorem 1. □

Remark 2.2 *In the proof of Theorem 1, we have obtained bounds on $|\lambda_\delta - \lambda|$ from (71), using L^p estimates on w_{u,u_δ} and $(u_\delta - u)$ to control the second term of the right hand side. Remarking that*

$$\begin{aligned}\nabla w_{u,u_\delta} &= -u \frac{f(u^2)u - f(u_\delta^2)u - 2f'(u_\delta^2)u_\delta^2(u - u_\delta)}{(u_\delta - u)^2} \nabla u_\delta \\ &\quad - u_\delta \frac{f(u_\delta^2)u_\delta - f(u^2)u_\delta - 2f'(u^2)u^2(u_\delta - u)}{(u_\delta - u)^2} \nabla u \\ &\quad + 2uu_\delta \left(f'(u_\delta^2) \nabla u_\delta + f'(u^2) \nabla u \right) + 2u_\delta \frac{f(u_\delta^2) - f(u^2)}{u_\delta - u} \nabla u_\delta \,,\end{aligned}$$

we can see that if u_δ is uniformly bounded in $L^\infty(\Omega)$ and if F satisfies (8) for $r = 2$ and is such that $F''(t)t^{1/2}$ is bounded in the vicinity of 0, then w_{u,u_δ} is uniformly bounded in X. It then follows from (71) that

$$|\lambda_\delta - \lambda| \leq C \left(\|u_\delta - u\|_{H^1}^2 + \|u_\delta - u\|_{X'} \right),$$

an estimate which is an improvement of (68). In the next two sections, we will see that this approach (or analogous strategies making use of negative Sobolev norms of higher orders), can be used in certain cases to obtain optimal estimates on $|\lambda_\delta - \lambda|$ of the form

$$|\lambda_\delta - \lambda| \leq C \|u_\delta - u\|_{H^1}^2,$$

similar to what is obtained for the linear eigenvalue problem $-\Delta u + Vu = \lambda u$.

3 Fourier expansion

In this section, we consider the problem

$$\inf\left\{E(v),\ v\in X,\ \int_\Omega v^2=1\right\}, \tag{40}$$

where

$$\begin{aligned}&\Omega=(0,2\pi)^d,\quad \text{with } d=1,\ 2 \text{ or } 3,\\ &X=H^1_\#(\Omega),\\ &E(v)=\frac{1}{2}\int_\Omega|\nabla v|^2+\frac{1}{2}\int_\Omega Vv^2+\frac{1}{2}\int_\Omega F(v^2).\end{aligned}$$

We assume that $V\in H^\sigma_\#(\Omega)$ for some $\sigma>d/2$ and that the function F satisfies (5)-(7), (8) for some $1<r\le 2$ and $0\le s\le 5-r$, and is in $C^{[\sigma]+1,\sigma-[\sigma]+\epsilon}((0,+\infty),\mathbb{R})$ (with the convention that $C^{k,0}=C^k$ if $k\in\mathbb{N}$).

The positive solution u to (40), which satisfies the elliptic equation

$$-\Delta u+Vu+f(u^2)u=\lambda u,$$

then is in $H^{\sigma+2}_\#(\Omega)$ and is bounded away from 0. To obtain this result, we have used the fact [47] that if $s>d/2$, $g\in C^{[s],s-[s]+\epsilon}(\mathbb{R},\mathbb{R})$ and $v\in H^s_\#(\Omega)$, then $g(v)\in H^s_\#(\Omega)$.

A natural discretization of (40) consists in using a Fourier basis. Denoting by $e_k(x)=(2\pi)^{-d/2}e^{ik\cdot x}$, we have for all $v\in L^2(\Omega)$,

$$v(x)=\sum_{\boldsymbol{k}\in\mathbb{Z}^d}\widehat{v}_k e_k(x),$$

where $\widehat{v}_k$ is the k^{th} Fourier coefficient of v:

$$\widehat{v}_k:=\int_\Omega v(x)\,\overline{e_k(x)}\,dx=(2\pi)^{-d/2}\int_\Omega v(x)\,e^{-ik\cdot x}\,dx.$$

The approximation of the solution to (40) by the spectral Fourier approximation is based on the choice

$$X_\delta=\widetilde{X}_N=\text{Span}\{e_k,\ |k|_*\le N\},$$

where $|k|_*$ denotes either the l^2 -norm or the l^∞ -norm of k (i.e. either $|k|=(\sum_{i=1}^d|k_i|^2)^{1/2}$ or $|k|_\infty=\max_{1\le i\le d}|k_i|$). For convenience, the discretization parameter for this approximation will be denoted as N.

Endowing $H^r_\#(\Omega)$ with the norm defined by

$$\|v\|_{H^r}=\left(\sum_{k\in\mathbb{Z}^d}\left(1+|k|_*^2\right)^r|\widehat{v}_k|^2\right)^{1/2},$$

we obtain that for all $s \in \mathbb{R}$, and all $v \in H^s_\#(\Omega)$, the best approximation of v in $H^r_\#(\Omega)$ for any $r \leq s$ is

$$\Pi_N v = \sum_{\boldsymbol{k}\in\mathbb{Z}^d, |\boldsymbol{k}|_* \leq N} \widehat{v}_k e_k.$$

The more regular v (the regularity being measured in terms of the Sobolev norms H^r), the faster the convergence of this truncated series to v: for all real numbers r and s with $r \leq s$, we have

$$\forall v \in H^s_\#(\Omega), \quad \|v - \Pi_N v\|_{H^r} \leq \frac{1}{N^{s-r}} \|v\|_{H^s}. \tag{41}$$

Let u_N be a solution to the variational problem

$$\inf \left\{ E(v_N),\ v_N \in \widetilde{X}_N,\ \int_\Omega v_N^2 = 1 \right\}$$

such that $(u_N, u)_{L^2} \geq 0$. Using (2), we obtain

$$\|u - \Pi_N u\|_{H^1} \leq \frac{1}{N^{\sigma+1}} \|u\|_{H^{\sigma+2}},$$

and it therefore follows from the first assertion of Theorem 1 that

$$\lim_{N\to\infty} \|u_N - u\|_{H^1} = 0.$$

We then observe that u_N is solution to the elliptic equation

$$-\Delta u_N + \Pi_N \left[V u_N + f(u_N^2) u_N \right] = \lambda_N u_N. \tag{42}$$

Thus u_N is uniformly bounded in $H^2_\#(\Omega)$, hence in $L^\infty(\Omega)$, and

$$\begin{aligned} \Delta (u_N - u) &= \Pi_N \left(V(u_N - u) + f(u_N^2) u_N - f(u^2) u \right) \\ &\quad -(I - \Pi_N)(Vu + f(u^2)u) - \lambda_N (u_N - u) - (\lambda_N - \lambda) u. \end{aligned} \tag{43}$$

As $(u_N)_{N\in\mathbb{N}}$ in bounded in $L^\infty(\Omega)$ and converges to u in $H^1_\#(\Omega)$, the right hand side of the above equality converges to 0 in $L^2_\#(\Omega)$, which implies that $(u_N)_{N\in\mathbb{N}}$ converges to 0 in $H^2_\#(\Omega)$, and therefore in $C^0_\#(\Omega)$. In particular, $u/2 \leq u_N \leq 2u$ on Ω for N large enough, so that we can assume in our analysis, without loss of generality, that F satisfies (6) with $q = 0$ and (8) with $r = 2$ and $s = 0$. We also deduce from (62) that u_N converges to u in $H^{\sigma+2}_\#(\Omega)$.

Besides, the unique solution to (64) solves the elliptic equation

$$\begin{aligned} -\Delta \psi_w + \left(V + f(u^2) + 2f'(u^2)u^2 - \lambda \right) \psi_w \\ = 2 \left(\int_\Omega f'(u^2) u^3 \psi_w \right) u + w - (w, u)_{L^2} u, \end{aligned} \tag{44}$$

from which we infer that $\psi_{u_N - u} \in H^2_\#(\Omega)$ and $\|\psi_{u_N-u}\|_{H^2} \leq C \|u_N - u\|_{L^2}$. Hence,

$$\|\psi_{u_N-u} - \Pi_N \psi_{u_N-u}\|_{H^1} \leq \frac{1}{N} \|\psi_{u_N-u}\|_{H^2} \leq \frac{C}{N} \|u_N - u\|_{L^2}.$$

We therefore deduce from Theorem 1 that

$$\begin{aligned} \|u_N - u\|_{H^s} &\leq \frac{C}{N^{\sigma+2-s}} \qquad \text{for } s = 0 \text{ and } s = 1 \quad (45) \\ |\lambda_N - \lambda| &\leq \frac{C}{N^{\sigma+2}} \quad (46) \\ \frac{\gamma}{2}\|u_N - u\|_{H^1}^2 \leq E(u_N) - E(u) &\leq C\|u_N - u\|_{H^1}^2. \end{aligned}$$

From (85) and the inverse inequality

$$\forall v_N \in \widetilde{X}_N, \quad \|v_N\|_{H^r} \leq 2^{(r-s)/2} N^{r-s} \|v_N\|_{H^s},$$

which holds true for all $s \leq r$ and all $N \geq 1$, we then obtain using classical arguments that

$$\|u_N - u\|_{H^s} \leq \frac{C}{N^{\sigma+2-s}} \qquad \text{for all } 0 \leq s < \sigma + 2. \quad (47)$$

The estimate (86) is slightly deceptive since, in the case of a linear eigenvalue problem (i.e. for $-\Delta u + Vu = \lambda u$) the convergence of the eigenvalues goes twice as fast as the convergence of the eigenvector in the H^1-norm. We are going to prove that this is also the case for the nonlinear eigenvalue problem under study in this section, at least under the assumption that $F \in C^{[\sigma]+2,\sigma-[\sigma]+\epsilon}((0,+\infty),\mathbb{R})$.

Let us first come back to (71), which we rewrite as,

$$\lambda_N - \lambda = \langle (A_u - \lambda)(u_N - u), (u_N - u)\rangle_{X',X} + \int_\Omega w_{u,u_N}(u_N - u) \quad (48)$$

with

$$w_{u,u_N} = u_N^2 \frac{f(u_N^2) - f(u^2)}{u_N - u} = u_N^2(u_N + u)\frac{f(u_N^2) - f(u^2)}{u_N^2 - u^2}.$$

As $u/2 \leq u_N \leq 2u$ on Ω for N large enough, as u_N converges, hence is uniformly bounded, in $H^{\sigma+2}_\#(\Omega)$ and as $f \in C^{[\sigma]+1,\sigma-[\sigma]+\epsilon}([\|u\|_{L^\infty}^2/4, 4\|u\|_{L^\infty}^2],\mathbb{R})$, we obtain that w_{u,u_N} is uniformly bounded in $H^\sigma_\#(\Omega)$ (at least for N large enough). We therefore infer from (89) that for N large enough

$$|\lambda_N - \lambda| \leq C\left(\|u_N - u\|_{H^1}^2 + \|u_N - u\|_{H^{-\sigma}}\right). \quad (49)$$

Let us now compute the H^{-r}-norm of the error for $0 < r \leq \sigma$. Let $w \in H^r_\#(\Omega)$. Proceeding as in Section 2, we obtain

$$\begin{aligned} \int_\Omega w(u_N - u) &= \langle (E''(u) - \lambda)(u_N - u), \Pi^1_{\widetilde{X}_N \cap u^\perp}\psi_w\rangle_{X',X} \\ &\quad + \langle (E''(u) - \lambda)(u_N - u), \psi_w - \Pi^1_{\widetilde{X}_N \cap u^\perp}\psi_w\rangle_{X',X} \\ &\quad + \|u_N - u\|_{L^2}^2 \int_\Omega f'(u^2)u^3\psi_w - \frac{1}{2}\|u_N - u\|_{L^2}^2 \int_\Omega uw, \quad (50) \end{aligned}$$

where $\Pi^1_{\widetilde{X}_N \cap u^\perp}$ denotes the orthogonal projector on $\widetilde{X}_N \cap u^\perp$ for the H^1 inner product. We then get from (44) that ψ_w is in $H^{r+2}_\#(\Omega)$ and satisfies

$$\|\psi_w\|_{H^{r+2}} \leq C\|w\|_{H^r}, \tag{51}$$

for some constant C independent of w.

Combining (18), (82), (84), (88), (89), (91) and (92), we obtain that there exists a constant $C \in \mathbb{R}_+$ such that for all $N \in \mathbb{N}$ and all $w \in H^r_\#(\Omega)$,

$$\begin{aligned}\int_\Omega w(u_N - u) &\leq C'\left(\|u_N - u\|^2_{L^2} + N^{-(r+1)}\|u_N - u\|_{H^1}\right)\|w\|_{H^r} \\ &\leq \frac{C}{N^{\sigma+2+r}}\|w\|_{H^r}.\end{aligned}$$

Therefore

$$\|u_N - u\|_{H^{-r}} = \sup_{w \in H^r_\#(\Omega)\setminus\{0\}} \frac{\displaystyle\int_\Omega w(u_N - u)}{\|w\|_{H^r}} \leq \frac{C}{N^{\sigma+2+r}}, \tag{52}$$

for some constant $C \in \mathbb{R}_+$ independent of N. Using (88) and (90), we end up with

$$|\lambda_N - \lambda| \leq \frac{C}{N^{2(\sigma+1)}}.$$

We can summarize the results obtained in this section in the following theorem.

Theorem 2 *Assume that $V \in H^\sigma_\#(\Omega)$ for some $\sigma > d/2$ and that the function F satisfies (5)-(7) and is in $C^{[\sigma]+1,\sigma-[\sigma]+\epsilon}((0,+\infty),\mathbb{R})$. Then $(u_N)_{N\in\mathbb{N}}$ converges to u in $H^{\sigma+2}_\#(\Omega)$ and there exists $C \in \mathbb{R}_+$ such that for all $N \in \mathbb{N}$,*

$$\|u_N - u\|_{H^s} \leq \frac{C}{N^{\sigma+2-s}} \qquad \text{for all } -\sigma \leq s < \sigma+2 \tag{53}$$

$$|\lambda_N - \lambda| \leq \frac{C}{N^{\sigma+2}}$$

$$\frac{\gamma}{2}\|u_N - u\|^2_{H^1} \leq E(u_N) - E(u) \leq C\|u_N - u\|^2_{H^1}. \tag{54}$$

If, in addition, $F \in C^{[\sigma]+2,\sigma-[\sigma]+\epsilon}((0,+\infty),\mathbb{R})$, then

$$|\lambda_N - \lambda| \leq \frac{C}{N^{2(\sigma+1)}}. \tag{55}$$

In order to evaluate the quality of the error bounds obtained in Theorem 2, we have performed numerical tests with $\Omega = (0, 2\pi)$, $V(x) = \sin(|x-\pi|/2)$ and $F(t^2) = t^2/2$. The Fourier coefficients of the potential V are given by

$$\widehat{V}_k = -\frac{1}{\sqrt{2\pi}}\frac{1}{|k|^2 - \frac{1}{4}}, \tag{56}$$

from which we deduce that $V \in H^{\sigma}_{\#}(0, 2\pi)$ for all $\sigma < 3/2$. It can be see on Figure 1 that $\|u_N - u\|_{H^1}$, $\|u_N - u\|_{L^2}$, $\|u_N - u\|_{H^{-1}}$, and $|\lambda_N - \lambda|$ decay respectively as $N^{-2.67}$, $N^{-3.67}$, $N^{-4.67}$ and N^{-5} (the reference values for u and λ are those obtained for $N = 65$). These results are in good agreement with the upper bounds (53) (for $s = 1$ and $s = 0$), (93) (for $r = 1$) and (55), which respectively decay as $N^{-2.5+\epsilon}$, $N^{-3.5+\epsilon}$, $N^{-4.5+\epsilon}$ and $N^{-5+\epsilon}$, for $\epsilon > 0$ arbitrarily small.

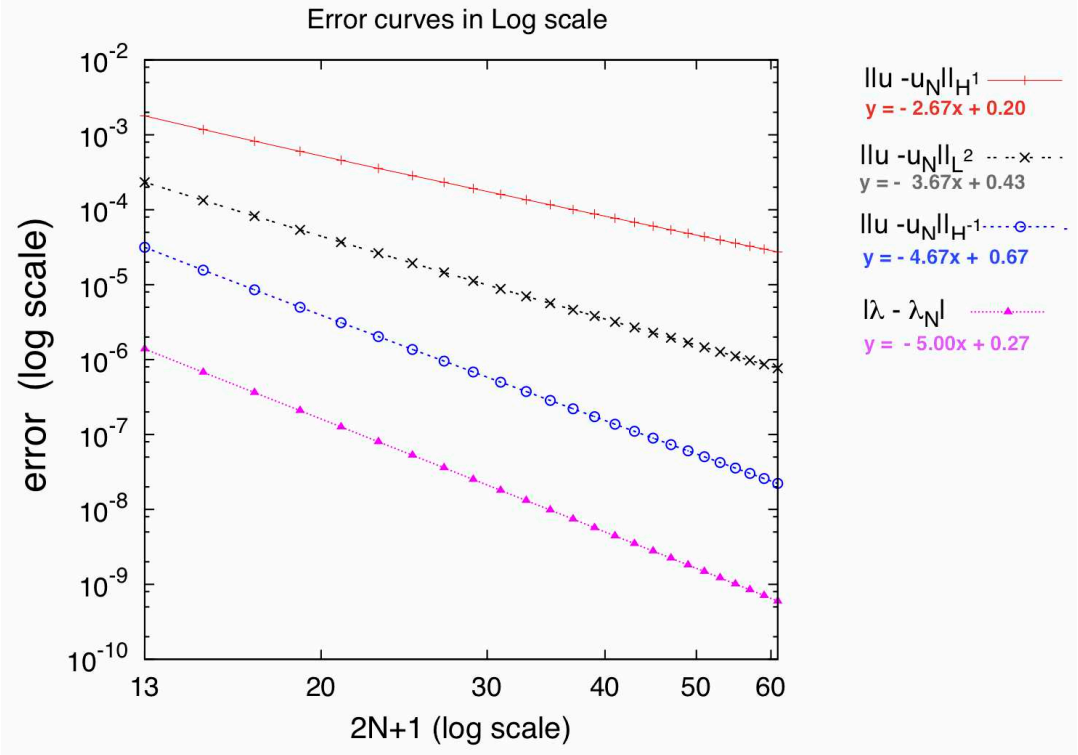

Figure 1: Numerical errors $\|u_N - u\|_{H^1}$, $\|u_N - u\|_{L^2}$, $\|u_N - u\|_{H^{-1}}$ and $|\lambda_N - \lambda|$ as functions of $2N + 1$ (the dimension of $\widetilde{X}_N$) in log scales.

4 Finite element discretization

In this section, we consider the problem

$$\inf \left\{ E(v),\ v \in X,\ \int_\Omega v^2 = 1 \right\}, \tag{57}$$

where

$$\begin{aligned} &\Omega \text{ is a rectangular brick of } \mathbb{R}^d, \quad \text{with } d = 1,\ 2 \text{ or } 3, \\ &X = H^1_0(\Omega), \\ &E(v) = \frac{1}{2}\int_\Omega |\nabla v|^2 + \frac{1}{2}\int_\Omega V v^2 + \frac{1}{2}\int_\Omega F(v^2). \end{aligned}$$

We assume that $V \in L^2(\Omega)$ and that the function F satisfies (5)-(7), as well as (8) for some $1 < r \leq 2$ and $0 \leq r + s \leq 3$. Throughout this section, we denote by u the unique positive solution of (57) and by λ the corresponding Lagrange multiplier.

In the non periodic case considered here, a classical variational approximation of (1) is provided by the finite element method. We consider a family of regular triangulations $(\mathcal{T}_h)_h$ of Ω. This means, in the case when $d = 3$ for instance, that for each $h > 0$, $\mathcal{T}_h$ is a collection of tetrahedra such that

- $\overline{\Omega}$ is the union of all the elements of $\mathcal{T}_h$;
- the intersection of two different elements of $\mathcal{T}_h$ is either empty, a vertex, a whole edge, or a whole face of both of them;
- the ratio of the diameter h_K of any element K of $\mathcal{T}_h$ to the diameter of its inscribed sphere is smaller than a constant independent of h.

As usual, h denotes the maximum of the diameters h_K, $K \in \mathcal{T}_h$. The parameter of the discretization then is $\delta = h > 0$. For each K in $\mathcal{T}_h$ and each nonnegative integer k, we denote by $\mathbb{P}_k(K)$ the space of the restrictions to K of the polynomials with d variables and total degree lower or equal to k.

The finite element space $X_{h,k}$ constructed from $\mathcal{T}_h$ and $\mathbb{P}_k(K)$ is the space of all continuous functions on Ω vanishing on $\partial\Omega$ such that their restrictions to any element K of $\mathcal{T}_h$ belong to $\mathbb{P}_k(K)$. Recall that $X_{h,k} \subset H^1_0(\Omega)$ as soon as $k \geq 1$.

We denote by $\pi^0_{h,k}$ and $\pi^1_{h,k}$ the orthogonal projectors on $X_{h,k}$ for the L^2 and H^1 inner products respectively. The following estimates are classical (see e.g. [15]): there exists $C \in \mathbb{R}_+$ such that for all $r \in \mathbb{N}$ such that $1 \leq r \leq k+1$,

$$\begin{aligned} &\forall \phi \in H^r(\Omega) \cap H^1_0(\Omega), \qquad \|\phi - \pi^0_{h,k}\phi\|_{L^2} \leq C h^r \|\phi\|_{H^r}, \\ &\forall \phi \in H^r(\Omega) \cap H^1_0(\Omega), \qquad \|\phi - \pi^1_{h,k}\phi\|_{H^1} \leq C h^{r-1} \|\phi\|_{H^r}. \end{aligned} \tag{58}$$

Let $u_{h,k}$ be a solution to the variational problem

$$\inf \left\{ E(v_{h,k}),\ v_{h,k} \in X_{h,k},\ \int_\Omega v^2_{h,k} = 1 \right\}$$

such that $(u_{h,k}, u)_{L^2} \geq 0$. In this setting, we obtain the following *a priori* error estimates.

Theorem 3 *Assume that $V \in L^2(\Omega)$ and that the function F satisfies (5), (6) for $q = 1$, (7), and (8) for some $1 < r \leq 2$ and $0 \leq r + s \leq 3$. Then there exists $h_0 > 0$ and $C \in \mathbb{R}_+$ such that for all $0 < h \leq h_0$,*

$$\|u_{h,1} - u\|_{H^1} \leq C\, h \tag{59}$$

$$\|u_{h,1} - u\|_{L^2} \leq C\, h^2 \tag{60}$$

$$|\lambda_{h,1} - \lambda| \leq C\, h^2 \tag{61}$$

$$\frac{\gamma}{2}\|u_{h,1} - u\|^2_{H^1} \leq E(u_{h,1}) - E(u) \leq C\, h^2. \tag{62}$$

If in addition, $V \in H^1(\Omega)$, F satisfies (8) for $r = 2$ and is such that $F \in C^3((0,+\infty),\mathbb{R})$ and $F''(t)t^{1/2}$ and $F'''(t)t^{3/2}$ are bounded in the vicinity of 0, then there exists $h_0 > 0$ and $C \in \mathbb{R}_+$ such that for all $0 < h \leq h_0$,

$$\begin{aligned} \|u_{h,2} - u\|_{H^1} &\leq C\,h^2 && (63)\\ \|u_{h,2} - u\|_{L^2} &\leq C\,h^3 && (64)\\ |\lambda_{h,2} - \lambda| &\leq C\,h^4 && (65)\\ \frac{\gamma}{2}\|u_{h,2} - u\|^2_{H^1} \leq E(u_{h,2}) - E(u) &\leq C\,h^4. && (66) \end{aligned}$$

Proof As Ω is a rectangular brick, V satisfies (4) and F satisfies (5)-(7), we have $u \in H^2(\Omega)$. We then use the fact that $\psi_{u_{h,k}-u}$ is solution to

$$\begin{aligned} &-\Delta\psi_{u_{h,k}-u} + (V + f(u^2) + 2f'(u^2)u^2 - \lambda)\psi_{u_{h,k}-u}\\ &\qquad = 2\left(\int_\Omega f'(u^2)u^3\psi_{u_{h,k}-u}\right)u + (u_{h,k} - u) - (u_{h,k} - u, u)_{L^2}u, \end{aligned}$$

to establish that $\psi_{u_{h,k}-u} \in H^2(\Omega) \cap H^1_0(\Omega)$ and that

$$\|\psi_{u_{h,k}-u}\|_{H^2} \leq C\|u_{h,k} - u\|_{L^2} \qquad (67)$$

for some constant C independent of h and k. The estimates (59)-(62) then are directly consequences of Theorem 1, (30), (58) and (67).

Under the additional assumptions that $V \in H^1(\Omega)$, we obtain by standard elliptic regularity arguments that $u \in H^3(\Omega)$. The H^1 and L^2 estimates (63) and (64) immediately follows from Theorem 1, (30), (58) and (67). We also have

$$|\lambda_{2,h} - \lambda| \leq Ch^3$$

for a constant C independent of h. In order to prove (65), we proceed as in Section 2. We start from the equality

$$\lambda_{2,h} - \lambda = \langle (A_u - \lambda)(u_{2,h} - u), (u_{2,h} - u)\rangle_{X',X} + \int_\Omega \widetilde{w}^h(u_{2,h} - u)$$

where

$$\widetilde{w}^h = u^2_{2,h}\frac{f(u^2_{2,h}) - f(u^2)}{u_{2,h} - u}.$$

We now claim that $u_{h,2}$ converges to u in $L^\infty(\Omega)$ when h goes to zero. To establish this result, we first remark that

$$\|u_{h,2} - u\|_{L^\infty} \leq \|u_{h,2} - \mathcal{I}_{h,2}u\|_{L^\infty} + \|\mathcal{I}_{h,2}u - u\|_{L^\infty},$$

where $\mathcal{I}_{h,2}$ is the interpolation projector on $X_{h,2}$. As $u \in H^3(\Omega) \hookrightarrow C^1(\overline{\Omega})$, we have

$$\lim_{h\to 0^+} \|\mathcal{I}_{h,2}u - u\|_{L^\infty} = 0.$$

On the other hand, using the inverse inequality

$$\exists C \in \mathbb{R}_+ \text{ s.t. } \forall 0 < h \leq h_0,\ \forall v_h \in X_{h,2}, \quad \|v_{h,2}\|_{L^\infty} \leq C\rho(h)\|v_{h,2}\|_{H^1},$$

with $\rho(h) = 1$ if $d = 1$, $\rho(h) = 1 + \ln h$ if $d = 2$ and $\rho(h) = h^{-1/2}$ if $d = 3$ (see [15] for instance), we obtain

$$\begin{aligned}\|u_{h,2} - \mathcal{I}_{h,2}u\|_{L^\infty} &\leq C\rho(h)\|u_{h,2} - \mathcal{I}_{h,2}u\|_{H^1} \\ &\leq C\rho(h)\left(\|u_{h,2} - u\|_{H^1} + \|u - \mathcal{I}_{h,2}u\|_{H^1}\right) \\ &\leq C'\,\rho(h)\,h^2 \underset{h\to 0^+}{\longrightarrow} 0.\end{aligned}$$

Hence the announced result. This implies in particular that $\widetilde{w}^h$ is bounded in $H^1(\Omega)$, uniformly in h. Consequently, there exists $C \in \mathbb{R}_+$ such that for all $0 < h \leq h_0$,

$$|\lambda_{h,2} - \lambda| \leq C\left(\|u_{h,2} - u\|_{H^1}^2 + \|u_{h,2} - u\|_{H^{-1}}\right). \tag{68}$$

To estimate the H^{-1}-norm of $u_{h,2} - u$, we write that for all $w \in H_0^1(\Omega)$,

$$\begin{aligned}\int_\Omega w(u_{h,2} - u) &= \langle (E''(u) - \lambda)(u_{h,2} - u), \pi^1_{X_{h,2}\cap u^\perp}\psi_w\rangle_{X',X} \\ &\quad + \langle (E''(u) - \lambda)(u_{h,2} - u), \psi_w - \pi^1_{X_{h,2}\cap u^\perp}\psi_w\rangle_{X',X} \\ &\quad + \|u_{h,2} - u\|_{L^2}^2 \int_\Omega f'(u^2)u^3\psi_w - \frac{1}{2}\|u_{h,2} - u\|_{L^2}^2 \int_\Omega uw,\end{aligned}$$

where ψ_w is solution to

$$\begin{aligned}&-\Delta\psi_w + (V + f(u^2) + 2f'(u^2)u^2 - \lambda)\psi_w \\ &\qquad = 2\left(\int_\Omega f'(u^2)u^3\psi_w\right)u + w - (w,u)_{L^2}u,\end{aligned} \tag{69}$$

and where $\pi^1_{X_{h,2}\cap u^\perp}$ denotes the orthogonal projector on $X_{h,2} \cap u^\perp$ for the H^1 inner product. Using the assumptions that $V \in H^1(\Omega)$, $F \in C^3((0,+\infty),\mathbb{R})$, and $F''(t)t^{1/2}$ and $F'''(t)t^{3/2}$ are bounded in the vicinity of 0, we deduce from (69) that ψ_w is in $H^3(\Omega)$ and that there exists $C \in \mathbb{R}_+$ such that for all $w \in H_0^1(\Omega)$ and all $0 < h \leq h_0$,

$$\|\psi_w\|_{H^3} \leq C\|w\|_{H^1}.$$

We therefore obtain the inequality

$$\|\psi_w - \pi^1_{h,2}\psi_w\|_{H^1} \leq Ch^2\|w\|_{H^1}, \tag{70}$$

where the constant C is independent of h.

Putting together (8) (for $r = 2$), (18), (82), (84), (58), (63), (64) and (70), we get

$$\|u_{h,2} - u\|_{H^{-1}} = \sup_{w\in H_0^1(\Omega)\setminus\{0\}} \frac{\int_\Omega w(u_{h,2} - u)}{\|w\|_{H^1}} \leq C\,h^4.$$

Combining with (63) and (68), we end up with (65). Lastly, we deduce (66) from the equality

$$\begin{aligned}E(u_{h,2}) - E(u) &= \frac{1}{2}\langle (A_u - \lambda)(u_{h,2} - u), (u_{h,2} - u)\rangle_{X',X} \\ &\quad + \frac{1}{2}\int_\Omega F\left(u^2 + (u_{h,2}^2 - u^2)\right) - F\left(u^2\right) - f\left(u^2\right)(u_{h,2}^2 - u^2),\end{aligned}$$

Taylor expanding the integrand and exploiting the boundedness of the function $F''(t)t^{1/2}$ in the vicinity of 0. □

Numerical results for the case when $\Omega = (0,\pi)^2$, $V(x_1,x_2) = x_1^2 + x_2^2$ and $F(t^2) = t^2/2$ are reported on Figure 2. The agreement with the error estimates obtained in Theorem 3 is good for the $\mathbb{P}_1$ approximation and excellent for the $\mathbb{P}_2$ approximation. Applied

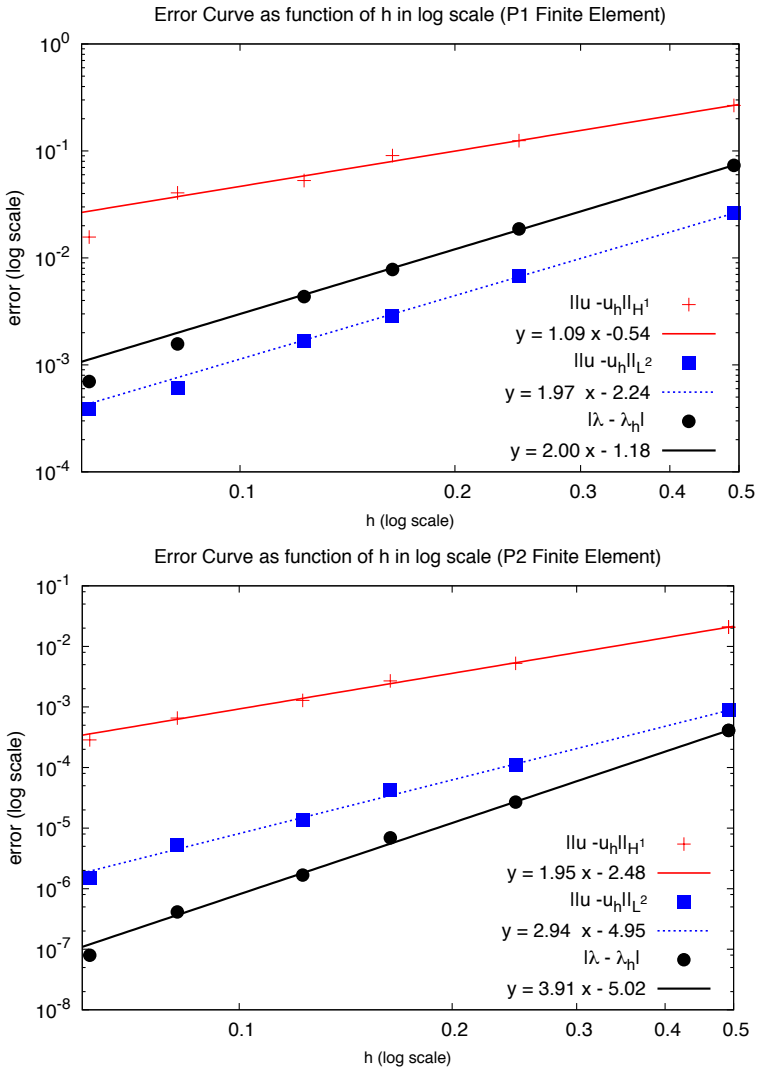

Figure 2: Errors $\|u_{h,k} - u\|_{H^1}$ (+), $\|u_{h,k} - u\|_{L^2}$ (■) and $|\lambda_{h,k} - \lambda|$ (•) for the $\mathbb{P}_1$ ($k = 1$, top) and $\mathbb{P}_2$ ($k = 2$, bottom) approximations as a function of h in log scales.

5 The effect of numerical integration

Let us now address one further consideration that is related to the practical implementation of the method, and more precisely to the numerical integration of the nonlinear term. For simplicity, we focus on the case when $A = 1$.

From a practical viewpoint, the solution $(u_\delta, \lambda_\delta)$ to the nonlinear eigenvalue problem (15) can be computed iteratively, using for instance the optimal damping algorithm [5, 6, 13]. At the p^{th} iteration ($p \geq 1$), the ground state $(u_\delta^p, \lambda_\delta^p) \in X_\delta \times \mathbb{R}$ of some *linear*, finite dimensional, eigenvalue problem of the form

$$\int_\Omega \overline{\nabla u_\delta^p} \cdot \nabla v_\delta + \int_\Omega \left(V + f(\widetilde{\rho}_\delta^{p-1})\right) \overline{u_\delta^p}\, v_\delta = \lambda_\delta^p \int_\Omega \overline{u_\delta^p} v_\delta, \quad \forall v_\delta \in X_\delta, \tag{71}$$

has to be computed. In the optimal damping algorithm, the density $\widetilde{\rho}_\delta^{p-1}$ is a convex linear combination of the densites $\rho_\delta^q = |u_\delta^q|^2$, for $0 \leq q \leq p-1$. Solving (71) amounts to finding the lowest eigenelement of the matrix H^p with entries

$$H_{kl}^p := \int_\Omega \overline{\nabla \phi_k} \cdot \nabla \phi_l + \int_\Omega V\, \overline{\phi_k}\, \phi_l + \int_\Omega f(\widetilde{\rho}_\delta^{p-1})\, \overline{\phi_k}\, \phi_l, \tag{72}$$

where $(\phi_k)_{1 \leq k \leq \dim(X_\delta)}$ stands for the canonical basis of X_δ.

In order to evaluate the last two terms of the right-hand side of (72), numerical integration has to be resorted to. In the finite element approximation of (57), it is generally made use of a numerical quadrature formula over each triangle (2D) or tetrahedron (3D) based on Gauss points. In the Fourier approximation of the periodic problem (40), the terms

$$\int_\Omega V\, \overline{e_k}\, e_l \quad \text{and} \quad \int_\Omega f(\widetilde{\rho}_\delta^{p-1})\, \overline{e_k}\, e_l,$$

which are in fact, up to a multiplicative constant, the $(k-l)^{\text{th}}$ Fourier coefficients of V and $f(\widetilde{\rho}_\delta^{p-1})$ respectively, are evaluated by Fast Fourier Transform (FFT), using an integration grid which may be different from the natural discretization grid

$$\left\{ \left(\frac{2\pi}{2N+1} j_1, \cdots, \frac{2\pi}{2N+1} j_d, \right),\ 0 \leq j_1, \cdots, j_d \leq 2N \right\}$$

associated with $\widetilde{X}_N$. This raises the question of the influence of the numerical integration on the convergence results obtained in Theorems 1, 2 and 3.

Remark 5.1 *In the case of the periodic problem considered in Section 2 and when $F(t) = ct^2$ for some $c > 0$, the last term of the right-hand side of (72) can be computed exactly (up to round-off errors) by means of a Fast Fourier Transform (FFT) on an integration grid twice as fine as the discretization grid. This is due to the fact that the function $\widetilde{\rho}_\delta^{p-1}\, \overline{e_k}\, e_l$ belongs to the space Span$\{e_n \,|\, |n|_* \leq 4N\}$. An analogous property is used in the evaluation of the Coulomb term in the numerical simulation of the Kohn-Sham equations for periodic systems.*

In the sequel, we focus on the simple case when $d = 1$, $\Omega = (0, 2\pi)$, $X = H^1_\#(0, 2\pi)$, and

$$E(v) = \frac{1}{2}\int_0^{2\pi} |v'|^2 + \frac{1}{2}\int_0^{2\pi} Vv^2 + \frac{1}{4}\int_0^{2\pi} |v|^4$$

with $V \in H^\sigma_\#(0, 2\pi)$ for some $\sigma > 1/2$. More difficult cases will be addressed elsewhere [8].

In view of Remark 5.1, we consider an integration grid

$$\frac{2\pi}{N_g}\mathbb{Z} \cap [0, 2\pi) = \left\{0, \frac{2\pi}{N_g}, \frac{4\pi}{N_g}, \cdots, \frac{2\pi(N_g - 1)}{N_g}\right\},$$

with $N_g \geq 4N + 1$ for which we have

$$\forall v_N \in \widetilde{X}_N, \quad \int_0^{2\pi} |v_N|^4 = \frac{2\pi}{N_g} \sum_{r \in \frac{2\pi}{N_g}\mathbb{Z} \cap [0,2\pi)} |v_N(r)|^4,$$

and for all $\rho \in \widetilde{X}_{2N}$,

$$\forall |k|, |l| \leq N, \quad \int_0^{2\pi} \rho\, \overline{e_k}\, e_l = \frac{1}{N_g} \sum_{r \in \frac{2\pi}{N_g}\mathbb{Z} \cap [0,2\pi)} \rho(r) e^{-i(k-l)r} = \widehat{\rho^{\mathrm{FFT}}_{k-l}}, \tag{73}$$

where $\widehat{\rho^{\mathrm{FFT}}_{k-l}}$ is the $(k-l)^{\mathrm{th}}$ coefficient of the discrete Fourier transform of ρ. Recall that if $\phi = \sum_{g\in\mathbb{Z}} \widehat{\phi}_g\, e_g \in C^0_\#(0, 2\pi)$, the discrete Fourier transform of ϕ is the $N_g\mathbb{Z}$-periodic sequence $(\widehat{\phi^{\mathrm{FFT}}_g})_{g\in\mathbb{Z}}$ defined by

$$\forall g \in \mathbb{Z}, \quad \widehat{\phi^{\mathrm{FFT}}_g} = \frac{1}{N_g} \sum_{r \in \frac{2\pi}{N_g}\mathbb{Z} \cap [0,2\pi)} \phi(r) e^{-igr}.$$

We now introduce the subspaces W_M for $M \in \mathbb{N}^*$ such that $W_M = \widetilde{X}_{(M-1)/2}$ if M is odd and $W_M = \widetilde{X}_{M/2-1} \oplus \mathbb{C}(e_{M/2} + e_{-M/2})$ is M is (note that $\dim(W_M) = M$ for all $M \in \mathbb{N}^*$). It is then possible to define an interpolation projector $\mathcal{I}_{N_g}$ from $C^0_\#(0, 2\pi)$ onto W_{N_g} by

$$\forall x \in \frac{2\pi}{N_g}\mathbb{Z} \cap [0, 2\pi), \quad [\mathcal{I}_{N_g}(\phi)](x) = \phi(x).$$

The expansion of $\mathcal{I}_{N_g}(\phi)$ in the canonical basis of W_{N_g} is given by

$$\mathcal{I}_{N_g}(\phi) = \left| \begin{array}{ll} (2\pi)^{1/2} \displaystyle\sum_{|g| \leq (N_g-1)/2} \widehat{\phi^{\mathrm{FFT}}_g}\, e_g & (N_g \text{ odd}), \\ (2\pi)^{1/2} \displaystyle\sum_{|g| \leq N_g/2-1} \widehat{\phi^{\mathrm{FFT}}_g}\, e_g + (2\pi)^{1/2} \widehat{\phi^{\mathrm{FFT}}_{N_g/2}} \left(\frac{e_{N_g/2} + e_{-N_g/2}}{2}\right) & (N_g \text{ even}). \end{array} \right.$$

Under the condition that $N_g \geq 4N+1$, the following property holds: for all $\phi \in C^0_\#(0,2\pi)$,

$$\forall |k|,|l| \leq N, \quad \int_0^{2\pi} \mathcal{I}_{N_g}(\phi)\,\overline{e_k}\,e_l = \widehat{\phi^{\mathrm{FFT}}_{k-l}}.$$

It is therefore possible, in the particular case considered here, to efficiently evaluate the entries of the matrix H^p using the formula

$$\begin{aligned} H^p_{kl} &:= \int_0^{2\pi} \overline{e'_k}\cdot e'_l + \int_0^{2\pi} V\,\overline{e_k}\,e_l + \int_0^{2\pi} \widetilde{\rho}_N^{p-1}\overline{e_k}\,e_l \\ &\simeq |k|^2\delta_{kl} + \widehat{V^{\mathrm{FFT}}_{k-l}} + \widehat{[\widetilde{\rho}_N^{p-1}]^{\mathrm{FFT}}_{k-l}}, \end{aligned} \tag{74}$$

and resorting to Fast Fourier Transform (FFT) algorithms to compute the discrete Fourier transforms. Note that only the second term is computed approximatively. The third term is computed exactly since, at each iteration, $\widetilde{\rho}_N^{p-1}$ belongs to $\widetilde{X}_{2N}$ (see Eq. (73)). Of course, this situation is specific to the nonlinearity $F(t) = t^2/2$ considered here.

Using the approximation formula (74) amounts to replace the original problem

$$\inf\left\{E(v_N),\ v_N \in \widetilde{X}_N,\ \int_0^{2\pi} |v_N|^2 = 1\right\}, \tag{75}$$

with the approximate problem

$$\inf\left\{E_{N_g}(v_N),\ v_N \in \widetilde{X}_N,\ \int_0^{2\pi} |v_N|^2 = 1\right\}, \tag{76}$$

where

$$E_{N_g}(v_N) = \frac{1}{2}\int_0^{2\pi} |v'_N|^2 + \frac{1}{2}\int_0^{2\pi} \mathcal{I}_{N_g}(V)v_N^2 + \frac{1}{4}\int_0^{2\pi} |v_N|^4.$$

Let us denote by u_N a solution of (75) such that $(u_N,u)_{L^2} \geq 0$ and by u_{N,N_g} a solution to (76) such that $(u_{N,N_g},u)_{L^2} \geq 0$. It is easy to check that u_{N,N_g} is bounded in $H^1_\#(0,2\pi)$ uniformly in N and N_g.

Besides, we know from Theorem 2 that $(u_N)_{N\in\mathbb{N}}$ converges to u in $H^1_\#(0,2\pi)$, hence in $L^\infty_\#(2,\pi)$, when N goes to infinity. This implies that the sequence $(A_u - A_{u_N})_{N\in\mathbb{N}}$ converges to 0 in operator norm. Consequently, for all N large enough and all N_g such that $N_g \geq 4N+1$,

$$\begin{aligned} \frac{\gamma}{4}\|u_{N,N_g} - u_N\|^2_{H^1} &\leq E(u_{N,N_g}) - E(u_N) \\ &\leq E_{N_g}(u_{N,N_g}) - E_{N_g}(u_N) \\ &\quad + \int_0^{2\pi} (V - \mathcal{I}_{N_g}(V))\left(|u_{N,N_g}|^2 - |u_N|^2\right) \\ &\leq \int_0^{2\pi} (V - \mathcal{I}_{N_g}(V))\left(|u_{N,N_g}|^2 - |u_N|^2\right) \\ &\leq C\|\Pi_{2N}(V - \mathcal{I}_{N_g}(V))\|_{L^2}\|u_{N,N_g} - u_N\|_{H^1}, \end{aligned}$$

where we have used the fact that $\left(|u_{N,N_g}|^2 - |u_N|^2\right) \in \widetilde{X}_{2N}$. Therefore,

$$\|u_{N,N_g} - u_N\|_{H^1} \leq C\|\Pi_{2N}(V - \mathcal{I}_{N_g}(V))\|_{L^2}, \tag{77}$$

for a constant C independent of N and N_g. Likewise,

$$\begin{aligned}\lambda_{N,N_g} - \lambda_N &= \langle (A_{u_N} - \lambda_N)(u_{N,N_g} - u_N), (u_{N,N_g} - u_N)\rangle_{X',X} \\ &\quad + \int_0^{2\pi} (V - \mathcal{I}_N(V))|u_{N,N_g}|^2 \\ &\quad + \int_0^{2\pi} |u_{N,N_g}|^2 (u_{N,N_g} + u_N)(u_{N,N_g} - u_N),\end{aligned}$$

from which we deduce, using (77),

$$|\lambda_{N,N_g} - \lambda_N| \leq C\|\Pi_{2N}(V - \mathcal{I}_{N_g}(V))\|_{L^2}.$$

An error analysis of the interpolation operator $\mathcal{I}_{N_g}$ is given in [10]: for all non-negative real numbers $0 \leq r \leq s$ with $s > 1/2$ (for $d = 1$),

$$\|\varphi - \mathcal{I}_{N_g}(\varphi)\|_{H^r} \leq \frac{C}{N_g^{s-r}}\|\varphi\|_{H^s}, \quad \forall \varphi \in H^s_{\#}(0, 2\pi).$$

Thus,

$$\|\Pi_{2N}(V - \mathcal{I}_{N_g}(V))\|_{L^2} \leq \|V - \mathcal{I}_{N_g}(V)\|_{L^2} \leq \frac{C}{N_g^{\sigma}}, \tag{78}$$

and the above inequality provides the following estimates:

$$\|u_{N,N_g} - u\|_{H^1} \leq C\left(N^{-\sigma-1} + N_g^{-\sigma}\right) \tag{79}$$

$$\|u_{N,N_g} - u\|_{L^2} \leq C\left(N^{-\sigma-2} + N_g^{-\sigma}\right) \tag{80}$$

$$|\lambda_{N,N_g} - \lambda| \leq C\left(N^{-2\sigma-2} + N_g^{-\sigma}\right), \tag{81}$$

for a constant C independent of N and N_g. The first component of the error bound (79) corresponds to the error $\|u_N - u\|_{H^1}$ while the second component corresponds to the numerical integration error $\|u_{N,N_g} - u_N\|_{H^1}$ (the same remark applies to the error bounds (80) and (81)).

It is classical that for the norm $\|\varphi - \mathcal{I}_{N_g}\varphi\|_{H^r}$ for $r < 0$ is in general of the same order of magnitude as $\|\varphi - \mathcal{I}_{N_g}\varphi\|_{L^2}$. As the existence of better estimates in negative norms is a corner stone in the derivation of the improvement of the error estimate (86) for the eigenvalues (doubling of the convergence rate), we expect that the eigenvalue approximation will be dramatically polluted by the use of the numerical integration formula.

This can be checked numerically. Considering again the one-dimensional example used in Section 2 ($\Omega = (0, 2\pi)$, $V(x) = \sin(|x - \pi|/2)$, $F(t) = t^2/2$), we have computed for $4 \leq N \leq 30$ and $N_g = 2^p$ with $7 \leq p \leq 15$, the errors $\|u_{N,N_g} - u\|_{H^1}$, $\|u_{N,N_g} - u\|_{L^2}$, $\|u_{N,N_g} - u\|_{H^{-1}}$, and $|\lambda_{N,N_g} - \lambda|$. On Figure 3, these quantities are plotted as functions of $2N + 1$ (the dimension of $\widetilde{X}_N$), for various values of N_g.

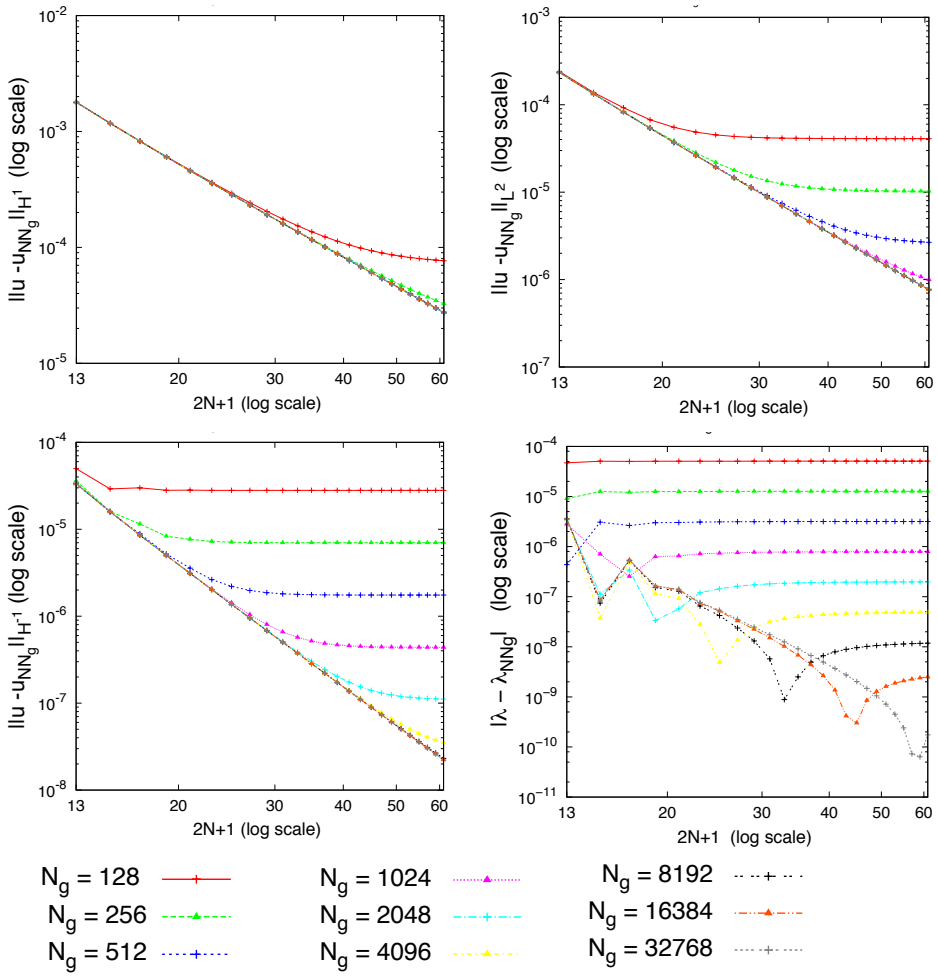

Figure 3: Numerical errors $\|u_{N,N_g} - u\|_{H^1}$ (top left), $\|u_{N,N_g} - u\|_{L^2}$ (top right), $\|u_{N,N_g} - u\|_{H^{-1}}$ (bottom left), and $|\lambda_{N,N_g} - \lambda|$ (bottom right), as functions of $2N+1$ (the dimension of $\widetilde{X}_N$) for $N_g = 2^p$, $7 \leq p \leq 15$.

The non-monotonicity of the curve $N \mapsto |\lambda_{N,N_g} - \lambda|$ originates from the fact that $\lambda_{N,N_g} - \lambda$ can be positive or negative depending on the values of N and N_g.

The numerical errors $\|u_{N,N_g} - u\|_{H^1}$, $\|u_{N,N_g} - u\|_{L^2}$, $\|u_{N,N_g} - u\|_{H^{-1}}$, and $|\lambda_{N,N_g} - \lambda|$, for $N = 30$, as functions of N_g (in log scales) are plotted on Figure 4. When N_g goes to infinity, the sequences $\log_{10}\|u_{N,N_g} - u\|_{H^1}$, $\log_{10}\|u_{N,N_g} - u\|_{L^2}$, $\log_{10}\|u_{N,N_g} - u\|_{H^{-1}}$, and $\log_{10}|\lambda_{N,N_g} - \lambda|$ converge to $\log_{10}\|u_N - u\|_{H^1}$, $\log_{10}\|u_N - u\|_{L^2}$, $\log_{10}\|u_N - u\|_{H^{-1}}$, and $\log_{10}|\lambda_N - \lambda|$ respectively. For smaller values of N_g, the numerical integration error dominates and these functions all decay linearly with $\log_{10} N_g$ with a slope very close to -2. For fixed N, the upper bounds (79)-(81) also decay linearly with $\log_{10} N_g$, but with a slope equal to -1.5. To obtain sharper upper bounds for the numerical integration error, we need to replace (78) with a sharper estimate of $\|\Pi_{2N}(V - \mathcal{I}_{N_g}(V))\|_{L^2}$,

which is possible for the particular example under consideration here. Indeed, remarking that under the condition $N_g \geq 4N+1$,

$$\|\Pi_{2N}(V - \mathcal{I}_{N_g}(V))\|_{L^2} = \left(\sum_{|g|\leq 2N} \left| \sum_{k\in\mathbb{Z}^*} \widehat{V}_{g+kN_g} \right|^2 \right)^{1/2},$$

we can, using (56), show that

$$\|\Pi_{2N}(V - \mathcal{I}_{N_g}(V))\|_{L^2} \leq \frac{C\, N^{1/2}}{N_g^2},$$

for a constant C independent of N and N_g. We deduce that for this specific example

$$\begin{aligned} \|u_{N,N_g} - u\|_{H^1} &\leq C\left(N^{-5/2} + N^{1/2}N_g^{-2}\right) \\ \|u_{N,N_g} - u\|_{L^2} &\leq C\left(N^{-7/2} + N^{1/2}N_g^{-2}\right) \\ |\lambda_{N,N_g} - \lambda| &\leq C\left(N^{-9/2} + N^{1/2}N_g^{-2}\right). \end{aligned}$$

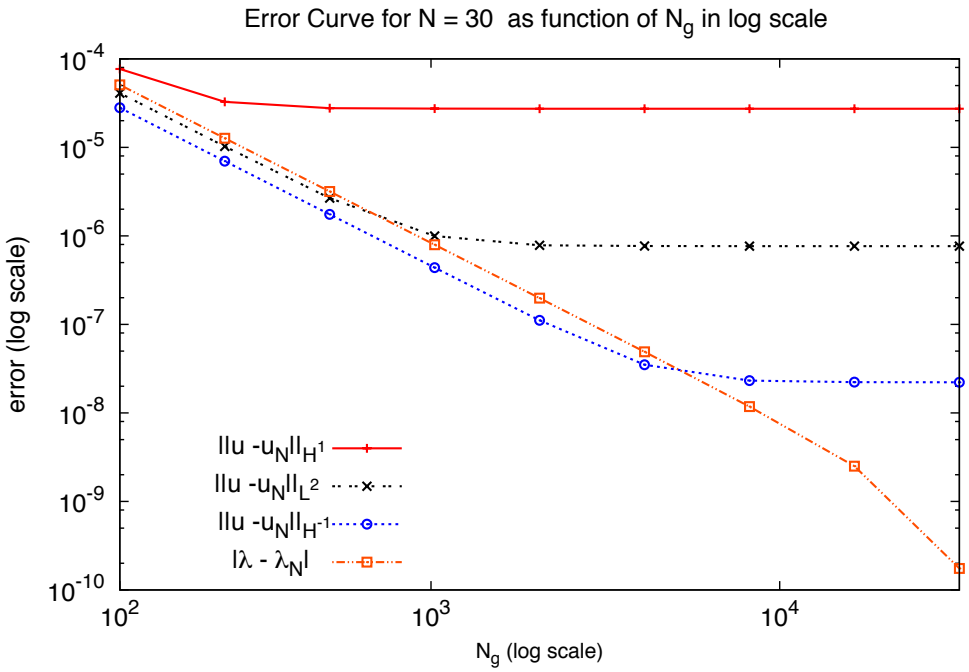

Figure 4: Numerical errors $\|u_{N,N_g} - u\|_{H^1}$ (+), $\|u_{N,N_g} - u\|_{L^2}$ (×), $\|u_{N,N_g} - u\|_{H^{-1}}$ (○), and $|\lambda_{N,N_g} - \lambda|$ (□), for $N = 30$, as functions of N_g (in log scales).

Acknowledgements

This work was done while E.C. was visiting the Division of Applied Mathematics of Brown University, whose support is gratefully acknowledged. The authors also thank Jean-Yves Chemin and Didier Smets for fruitful discussions, and Claude Le Bris for valuable comments on a preliminary version of this work.

6 Appendix: properties of the ground state

The mathematical properties of the minimization problems (1) and (9) which are useful for the numerical analysis reported in this article are gathered in the following lemma.

Recall that $d = 1$, 2 or 3.

Lemma 2 *Under assumptions (2)-(6), (9) has a unique minimizer ρ_0 and (1) has exactly two minimizers $u = \sqrt{\rho_0}$ and $-u$. The function u is solution to the nonlinear eigenvalue problem (11) for some $\lambda \in \mathbb{R}$. Besides, $u \in C^{0,\alpha}(\overline{\Omega})$ for some $0 < \alpha < 1$, $u > 0$ in Ω, and λ is the lowest eigenvalue of A_u and is non-degenerate.*

Proof As A is uniformly bounded and coercive on Ω and $V \in L^q(\Omega)$ for some $q > \max(1, d/2)$, $v \mapsto a(v,v)$ is a quadratic form on X, bounded from below on the set $\{v \in X \mid \|v\|_{L^2} = 1\}$. Replacing $a(v,v)$ with $a(v,v) + C\|v\|_{L^2}^2$ and $F(t)$ with $F(t) - F(0) - tF'(0)$ does not change the minimizers of (1) and (9). We can therefore assume, *without loss of generality*, that

$$\forall v \in X,\ a(v,v) \geq \|v\|_{L^2}^2 \quad \text{and} \quad F(0) = F'(0) = 0. \tag{82}$$

It then follows from (6) and (82) that $0 \leq F(v^2) \leq C(v^2 + v^6)$. As $X \hookrightarrow L^6(\Omega)$, $E(v)$ is finite for all $v \in X$, $I > -\infty$ and the minimizing sequences of (1) are bounded in X. Let $(v_n)_{n\in\mathbb{N}}$ be a minimizing sequence of (1). Using the fact that X is compactly embedded in $L^2(\Omega)$, we can extract from $(v_n)_{n\in\mathbb{N}}$ a subsequence $(v_{n_k})_{k\in\mathbb{N}}$ which converges weakly in X, strongly in $L^2(\Omega)$ and almost everywhere in Ω to some $u \in X$. As $\|v_{n_k}\|_{L^2} = 1$ and $E(v_{n_k}) \downarrow I$, we obtain $\|u\|_{L^2} = 1$ and $E(u) \leq I$ (E is convex and strongly continuous, hence weakly l.s.c., on X). Hence u is a minimizer of (1). As $|u| \in X$, $\||u|\|_{L^2} = 1$ and $E(|u|) = E(u)$, we can assume without loss of generality that $u \geq 0$. Assumptions (2)-(6) imply that E is C^1 on X and that $E'(u) = A_u u$. It follows that u is solution to (10) for some $\lambda \in \mathbb{R}$. By elliptic regularity arguments [19], we get $u \in C^{0,\alpha}(\overline{\Omega})$ for some $0 < \alpha < 1$. We also have $u > 0$ in Ω; this is a consequence of the Harnack inequality [49]. Making the change of variable $\rho = v^2$, it is easily seen that if v is a minimizer of (1), then v^2 is a minimizer of (9), and that, conversely, if ρ is a minimizer of (9), then $\sqrt{\rho}$ and $-\sqrt{\rho}$ are minimizers of (1). Besides, the

functional $\mathcal{E}$ is strictly convex on the convex set $\left\{\rho \geq 0 \mid \sqrt{\rho} \in X,\ \int_\Omega \rho = 1\right\}$. Therefore $\rho_0 = u^2$ is the unique minimizer of (9) and u and $-u$ are the only minimizers of (1).

It is easy to see that A_u is bounded below and has a compact resolvent. It therefore possesses a lowest eigenvalue λ_0, which, according to the min-max principle, satisfies

$$\lambda_0 = \inf\left\{\int_\Omega (A\nabla v)\cdot\nabla v + \int_\Omega (V + f(u^2))v^2,\ v \in X,\ \int_\Omega v^2 = 1\right\}. \tag{83}$$

Let v_0 be a normalized eigenvector of A_u associated with λ_0. Clearly, v_0 is a minimizer of (83) and so is $|v_0|$. Therefore, $|v_0|$ is solution to the Euler equation $A_u|v_0| = \lambda_0|v_0|$. Using again elliptic regularity arguments and the Harnack inequality, we obtain that $|v_0| \in C^{0,\alpha}(\overline{\Omega})$ for some $0 < \alpha < 1$ and that $|v_0| > 0$ on Ω. This implies that either $v_0 = |v_0| > 0$ in Ω or $v_0 = -|v_0| < 0$ in Ω. In particular $(u, v_0)_{L^2} \neq 0$. Consequently, $\lambda = \lambda_0$ and λ is a simple eigenvalue of A_u. □

Let us finally prove that λ is also the ground state eigenvalue of the *nonlinear* eigenvalue problem

$$\begin{cases} \text{search } (\mu, v) \in \mathbb{R}\times X \text{ such that} \\ A_v v = \mu v \\ \|v\|_{L^2} = 1, \end{cases} \tag{84}$$

in the following sense: if (μ, v) is solution to (84) then either $\mu > \lambda$ or $\mu = \lambda$ and $v = \pm u$.

To see this, let us consider a solution $(\mu, v) \in \mathbb{R}\times X$ to (84) and denote by $\widetilde{w} = |v| - u$. As for u, we infer from elliptic regularity arguments [19] that $v \in C^{0,\alpha}(\overline{\Omega})$. We have $\|v\|_{L^2} = \|u\|_{L^2} = 1$. Therefore, if $w \leq 0$ in Ω, then $|v| = u$, which yields $v = \pm u$ and $\mu = \lambda$. Otherwise, there exists $x_0 \in \Omega$ such that $\widetilde{w}(x_0) > 0$, and, up to replacing v with $-v$, we can consider that the function $w = v - u$ is such that $w(x_0) > 0$. The function w is in $X \cap C^{0,\alpha}(\overline{\Omega})$ and satisfies

$$(A_u - \lambda)w + \frac{f(v^2) - f(u^2)}{v^2 - u^2}v(u+v)w = (\mu - \lambda)v. \tag{85}$$

Let $\omega = \{x \in \Omega \mid w(x) > 0\} = \{x \in \Omega \mid v(x) > u(x)\}$ and $w_+ = \max(w, 0)$. As $w_+ \in X$, we deduce from (85) that

$$\langle (A_u - \lambda)w_+, w_+\rangle_{X',X} + \int_\omega \frac{f(v^2) - f(u^2)}{v^2 - u^2}v(u+v)w^2 = (\mu - \lambda)\int_\omega vw.$$

The left hand side of the above equality is positive and $\int_\omega vw > 0$. Therefore, $\mu > \lambda$.

Chapitre 2

Analyse numérique de problèmes aux valeurs propres non linéaire : le modèle de Thomas-Fermi-von-Weizäcker (TFW)

1 Introduction

Density Functional Theory (DFT) is a powerful method for computing ground state electronic energies and densities in quantum chemistry, materials science, molecular biology and nanosciences. The models originating from DFT can be classified into two categories: the orbital-free models and the Kohn-Sham models. The Thomas-Fermi-von Weizsäcker (TFW) model falls into the first category. It is not very much used in practice, but is interesting from a mathematical viewpoint. It indeed serves as a toy model for the analysis of the more complex electronic structure models routinely used by Physicists and Chemists. At the other extremity of the spectrum, the Kohn-Sham models are among the most widely used models in Physics and Chemistry, but are much more difficult to deal with. We focus here on the numerical analysis of the TFW model, more precisely, we are interested in the pseudospectral Fourier, more commonly called planewave, discretization of the periodic version of this model. In this context, the simulation domain, sometimes refered to as the supercell, is the unit cell of some periodic lattice of $\mathbb{R}^3$. Imposing periodic boundary (PBC) to the density at the boundary of the simulation cell is a standard method to compute condensed phase properties with a limited number of atoms in the simulation cell, hence at a moderate computational cost.

This chapter is organized as follows. In Section 2, we briefly recall the functional setting used in the formulation and the analysis of the planewave discretization of orbital-free previously introduced in chapter 1. In Section 3, we provide *a priori* error estimates for the planewave discretization of the TFW model. Our estimates refine and complement some of the results given in [28,59].

2 Basic Fourier analysis for planewave discretization methods

Throughout this chapter, we denote by Γ the simulation cell, by $\mathcal{R}$ the periodic lattice, and by $\mathcal{R}^*$ the dual lattice. For simplicity, we assume that $\Gamma = [0, L)^3$ ($L > 0$), but our arguments can be straightforwardly extended to rectangular simulation cells ($\Gamma = [0, L_x) \times [0, L_y) \times [0, L_z)$). For $\Gamma = [0, L)^3$, $\mathcal{R}$ is the cubic lattice $L\mathbb{Z}^3$, and $\mathcal{R}^* = \frac{2\pi}{L}\mathbb{Z}^3$. For $k \in \mathcal{R}^*$, we denote by $e_k(x) = |\Gamma|^{-1/2}\, e^{ik\cdot x}$ the planewave with wavevector k. The family $(e_k)_{k\in\mathcal{R}^*}$ forms an orthonormal basis of

$$L^2_{\#}(\Gamma, \mathbb{C}) := \left\{ u \in L^2_{\mathrm{loc}}(\mathbb{R}^3, \mathbb{C}) \mid u \ \mathcal{R}\text{-periodic} \right\},$$

and for all $u \in L^2_{\#}(\Gamma, \mathbb{C})$,

$$u(x) = \sum_{k\in\mathcal{R}^*} \widehat{u}_k\, e_k(x) \qquad \text{with} \qquad \widehat{u}_k = (e_k, u)_{L^2_{\#}} = |\Gamma|^{-1/2} \int_\Gamma u(x) e^{-ik\cdot x}\, dx.$$

In our analysis, we will only consider real valued functions. We therefore introduce the Sobolev spaces of real valued functions

$$H^s_\#(\Gamma) := \left\{ u(x) = \sum_{k\in\mathcal{R}^*} \widehat{u}_k\, e_k(x) \mid \sum_{k\in\mathcal{R}^*} (1+|k|^2)^s |\widehat{u}_k|^2 < \infty \text{ and } \forall k,\ c_{-k} = \overline{c}_k \right\},$$

$s \in \mathbb{R}$, endowed with the inner products

$$(u,v)_{H^s_\#} = \sum_{k\in\mathcal{R}^*} (1+|k|^2)^s\, \overline{\widehat{u}}_k\, \widehat{v}_k.$$

For $N_c \in \mathbb{N}$, we denote by

$$V_{N_c} = \left\{ \sum_{k\in\mathcal{R}^* \,|\, |k|\le \frac{2\pi}{L}N_c} c_k e_k \mid \forall k,\ c_{-k} = \overline{c}_k \right\} \tag{1}$$

(the constraints $c_{-k} = \overline{c}_k$ imply that the functions of V_{N_c} are real valued). For all $s \in \mathbb{R}$, and each $v \in H^s_\#(\Gamma)$, the best approximation of v in V_{N_c} for *any* $H^r_\#$-norm, $r \le s$, is

$$\Pi_{N_c} v = \sum_{k\in\mathcal{R}^* \,|\, |k|\le \frac{2\pi}{L}N_c} \widehat{v}_k e_k.$$

The more regular v (the regularity being measured in terms of the Sobolev norms H^r), the faster the convergence of this truncated series to v: for all real numbers r and s with $r \le s$, we have for each $v \in H^s_\#(\Gamma)$,

$$\begin{aligned} \|v - \Pi_{N_c} v\|_{H^r_\#} = \min_{v_{N_c}\in V_{N_c}} \|v - v_{N_c}\|_{H^r_\#} &\le \left(\frac{L}{2\pi}\right)^{s-r} N_c^{-(s-r)} \|v - \Pi_{N_c} v\|_{H^s_\#} \\ &\le \left(\frac{L}{2\pi}\right)^{s-r} N_c^{-(s-r)} \|v\|_{H^s_\#}. \end{aligned} \tag{2}$$

For $N_g \in \mathbb{N}\setminus\{0\}$, we denote by $\widehat{\phi}^{\mathrm{FFT},N_g}$ the discrete Fourier transform on the cartesian grid $\mathcal{G}_{N_g} := \frac{L}{N_g}\mathbb{Z}^3$ of the function $\phi \in C^0_\#(\Gamma)$. Recall that if $\phi = \sum_{k\in\mathcal{R}^*} \widehat{\phi}_g\, e_g \in C^0_\#(\Gamma)$, the discrete Fourier transform of ϕ is the $N_g\mathcal{R}^*$-periodic sequence $\widehat{\phi}^{\mathrm{FFT},N_g} = (\widehat{\phi}^{\mathrm{FFT},N_g}_k)_{k\in\mathcal{R}^*}$ where

$$\widehat{\phi}^{\mathrm{FFT},N_g}_k = \frac{1}{N_g^3} \sum_{x\in\mathcal{G}_{N_g}\cap\Gamma} \phi(x) e^{-ik\cdot x} = |\Gamma|^{-1/2} \sum_{K\in\mathcal{R}^*} \widehat{\phi}_{k+N_gK}.$$

We now introduce the subspaces

$$W^{1\mathrm{D}}_{N_g} = \left| \begin{array}{ll} \mathrm{Span}\left\{ e^{ily} \mid l \in \dfrac{2\pi}{L}\mathbb{Z},\ |l| \le \dfrac{2\pi}{L}\left(\dfrac{N_g-1}{2}\right)\right\} & (N_g \text{ odd}), \\ \mathrm{Span}\left\{ e^{ily} \mid l \in \dfrac{2\pi}{L}\mathbb{Z},\ |l| \le \dfrac{2\pi}{L}\left(\dfrac{N_g}{2}\right)\right\} \oplus \mathbb{C}(e^{i\pi N_g y/L} + e^{-i\pi N_g y/L}) & (N_g \text{ even}), \end{array} \right.$$

($W_{N_g}^{\rm 1D} \in C^\infty_\#([0,L))$ and $\dim(W_{N_g}^{\rm 1D}) = N_g$), and $W_{N_g}^{\rm 3D} = W_{N_g}^{\rm 1D} \otimes W_{N_g}^{\rm 1D} \otimes W_{N_g}^{\rm 1D}$. Note that $W_{N_g}^{\rm 3D}$ is a subspace of $H^s_\#(\Gamma)$ of dimension N_g^3, for all $s \in \mathbb{R}$, and that if N_g is odd,

$$W_{N_g}^{\rm 3D} = {\rm Span}\left\{ e_k \mid k \in \mathcal{R}^* = \frac{2\pi}{L}\mathbb{Z}^3,\ |k|_\infty \leq \frac{2\pi}{L}\left(\frac{N_g-1}{2}\right)\right\} \qquad (N_g \text{ odd}).$$

It is then possible to define the interpolation projector $\mathcal{I}_{N_g}$ from $C^0_\#(\Gamma)$ onto $W_{N_g}^{\rm 3D}$ by $[\mathcal{I}_{N_g}(\phi)](x) = \phi(x)$ for all $x \in \mathcal{G}_{N_g}$. It holds

$$\forall \phi \in C^0_\#(\Gamma), \quad \int_\Gamma \mathcal{I}_{N_g}(\phi) = \sum_{x \in \mathcal{G}_{N_g} \cap \Gamma} \left(\frac{L}{N_g}\right)^3 \phi(x). \tag{3}$$

The coefficients of the expansion of $\mathcal{I}_{N_g}(\phi)$ in the canonical basis of $W_{N_g}^{\rm 3D}$ is given by the discrete Fourier transform of ϕ. In particular, when N_g is odd, we have the simple relation

$$\mathcal{I}_{N_g}(\phi) = |\Gamma|^{1/2} \sum_{k \in \mathcal{R}^* \,|\, |k|_\infty \leq \frac{2\pi}{L}\left(\frac{N_g-1}{2}\right)} \widehat{\phi}_k^{{\rm FFT},N_g}\, e_k \qquad (N_g \text{ odd}).$$

It is easy to check that if ϕ is real-valued, then so is $\mathcal{I}_{N_g}(\phi)$.

We will assume in the sequel that $N_g \geq 4N_c + 1$. We will then have for all $v_{4N_c} \in V_{4N_c}$,

$$\int_\Gamma v_{4N_c} = \sum_{x \in \mathcal{G}_{N_g} \cap \Gamma} \left(\frac{L}{N_g}\right)^3 v_{4N_c}(x) = \int_\Gamma \mathcal{I}_{N_g}(v_{4N_c}). \tag{4}$$

The following lemma gathers some technical results which will be useful for the numerical analysis of the planewave discretization of orbital-free and Kohn-Sham models.

Lemma 2.1 *Let $N_c \in \mathbb{N}^*$ and $N_g \in \mathbb{N}^*$ such that $N_g \geq 4N_c + 1$.*

1. *Let V be a real-valued function of $C^0_\#(\Gamma)$ and v_{N_c} and w_{N_c} be two functions of V_{N_c}. Then*

$$\int_\Gamma \mathcal{I}_{N_g}(V v_{N_c} w_{N_c}) = \int_\Gamma \mathcal{I}_{N_g}(V) v_{N_c} w_{N_c} \tag{5}$$

$$\left|\int_\Gamma \mathcal{I}_{N_g}(V|v_{N_c}|^2)\right| \leq \|V\|_{L^\infty} \|v_{N_c}\|^2_{L^2_\#} \tag{6}$$

2. *Let $s > 3/2$, $0 \leq r \leq s$, and V a function of $H^s_\#(\Gamma)$. Then,*

$$\left\|(1 - \mathcal{I}_{N_g})(V)\right\|_{H^r_\#} \leq C_{r,s} N_g^{r-s} \|V\|_{H^s_\#} \tag{7}$$

$$\left\|\Pi_{2N_c}(\mathcal{I}_{N_g}(V))\right\|_{L^2_\#} \leq \left(\int_\Gamma \mathcal{I}_{N_g}(|V|^2)\right)^{1/2} \tag{8}$$

$$\left\|\Pi_{2N_c}(\mathcal{I}_{N_g}(V))\right\|_{H^s_\#} \leq (1 + C_{s,s}) \|V\|_{H^s_\#} \tag{9}$$

for constants $C_{r,s}$ independent of V. Besides if there exists $m > 3$ and $C \in \mathbb{R}_+$ such that $|\widehat{V}_k| \leq C|k|^{-m}$, then there exists a constant C_V independent of N_c and N_g such that

$$\left\|\Pi_{2N_c}(1-\mathcal{I}_{N_g})(V)\right\|_{H^r} \leq C_V N_c^{r+3/2} N_g^{-m} \tag{10}$$

3. *Let ϕ is be a continuous function from $\mathbb{R}_+$ to $\mathbb{R}$ such that there exists $C_\phi \in \mathbb{R}_+$ for which $|\phi(t)| \leq C_\phi(1+t^2)$ for all $t \in \mathbb{R}_+$. Then, for all $v_{N_c} \in V_{N_c}$,*

$$\left|\int_\Gamma \mathcal{I}_{N_g}(\phi(|v_{N_c}|^2))\right| \leq C_\phi\left(|\Gamma| + \|v_{N_c}\|_{L^4_\#}^4\right). \tag{11}$$

Proof For $z_{2N_c} \in V_{2N_c}$, it holds

$$\begin{aligned}\int_\Gamma \mathcal{I}_{N_g}(V z_{2N_c}) &= \sum_{x\in\mathcal{G}_{N_g}\cap\Gamma} V(x) z_{2N_c}(x) \\ &= \sum_{x\in\mathcal{G}_{N_g}\cap\Gamma} (\mathcal{I}_{N_g}(V))(x) z_{2N_c}(x) \\ &= \int_\Gamma \mathcal{I}_{N_g}(V)\, z_{2N_c} \end{aligned} \tag{12}$$

since $\mathcal{I}_{N_g}(V) z_{2N_c} \in V_{4N_c}$. The function $v_{N_c} w_{N_c}$ being in V_{2N_c}, (5) is proved. Moreover, as $|v_{N_c}|^2 \in V_{4N_c}$, it follows from (4) that

$$\begin{aligned}\left|\int_\Gamma \mathcal{I}_{N_g}(V|v_{N_c}|^2)\right| &= \left|\sum_{x\in\mathcal{G}_{N_g}\cap\Gamma} \left(\frac{L}{N_g}\right)^3 V(x)|v_{N_c}(x)|^2\right| \\ &\leq \|V\|_{L^\infty}\left|\sum_{x\in\mathcal{G}_{N_g}\cap\Gamma} \left(\frac{L}{N_g}\right)^3 |v_{N_c}(x)|^2\right| \\ &= \|V\|_{L^\infty}\int_\Gamma |v_{N_c}|^2. \end{aligned}$$

Hence (6). The estimate (7) is proved in [10]. To prove (8), we notice that

$$\begin{aligned}\|\Pi_{2N_c}(I_{N_g}(V))\|_{L^2_\#}^2 &\leq \|I_{N_g}(V)\|_{L^2_\#}^2 \\ &= \int_\Gamma (\overline{I_{N_g}(V)})(I_{N_g}(V)) \\ &= \sum_{x\in\mathcal{G}_{N_g}\cap\Gamma} \overline{(I_{N_g}(V))(x)}(I_{N_g}(V))(x) \\ &= \sum_{x\in\mathcal{G}_{N_g}\cap\Gamma} |V(x)|^2 \\ &= \int_\Gamma I_{N_g}(|V|^2). \end{aligned}$$

The bound (9) is a straightforward consequence of (7):

$$\|\Pi_{2N_c}(I_{N_g}(V))\|_{H^s_\#} \leq \|I_{N_g}(V)\|_{H^s_\#} \leq \|V\|_{H^s_\#} + \|(1-I_{N_g})(V)\|_{H^s_\#} \leq (1+C_{s,s})\|V\|_{H^s_\#}.$$

Now, we notice that

$$\begin{aligned}\Pi_{2N_c}(\mathcal{I}_{N_g}(V)) &= |\Gamma|^{1/2} \sum_{k\in\mathcal{R}^* \,|\, |k|\leq \frac{4\pi}{L}N_c} \widehat{V}_k^{\mathrm{FFT},N_g} e_k \\ &= \sum_{k\in\mathcal{R}^* \,|\, |k|\leq \frac{4\pi}{L}N_c} \left(\sum_{K\in\mathcal{R}^*} \widehat{V}_{k+N_gK} \right) e_k. \qquad (13)\end{aligned}$$

From (13), we obtain

$$\begin{aligned}\left\|\Pi_{2N_c}(1-\mathcal{I}_{N_g})(V)\right\|^2_{H^s} &= \sum_{k\in\mathcal{R}^* \,|\, |k|\leq \frac{4\pi}{L}N_c} (1+|k|^2)^s \left| \sum_{K\in\mathcal{R}^*\setminus\{0\}} \widehat{V}_{k+N_gK} \right|^2 \\ &= \left(\sum_{k\in\mathcal{R}^* \,|\, |k|\leq \frac{4\pi}{L}N_c} (1+|k|^2)^s \right) \max_{k\in\mathcal{R}^* \,|\, |k|\leq \frac{4\pi}{L}N_c} \left| \sum_{K\in\mathcal{R}^*\setminus\{0\}} \widehat{V}_{k+N_gK} \right|^2.\end{aligned}$$

On the one hand,

$$\sum_{k\in\mathcal{R}^* \,|\, |k|\leq \frac{4\pi}{L}N_c} (1+|k|^2)^s \underset{N_c\to\infty}{\sim} \frac{32\pi}{2s+3} \left(\frac{4\pi}{L}\right)^{2s} N_c^{2s+3},$$

and on the other hand, we have for each $k \in \mathcal{R}^*$ such that $|k| \leq \frac{4\pi}{L}N_c$,

$$\begin{aligned}\left| \sum_{K\in\mathcal{R}^*\setminus\{0\}} \widehat{V}_{k+N_gK} \right| &\leq C \sum_{K\in\mathcal{R}^*\setminus\{0\}} \frac{1}{|k+N_gK|^m} \\ &\leq C_0 \left(\frac{L}{2\pi}\right)^m N_g^{-m}\end{aligned}$$

where

$$C_0 = \max_{y\in\mathbb{R}^3 \,|\, |y|\leq 1/2} \sum_{K\in\mathbb{Z}^3\setminus\{0\}} \frac{1}{|y-K|^m}.$$

The estimate (10) then easily follows. Let us finally prove (11). Using (3) and (4), we have

$$\begin{aligned}\left| \int_\Gamma \mathcal{I}_{N_g}(\phi(|v_N|^2)) \right| &= \left| \sum_{x\in\frac{L}{N_g}\mathbb{Z}^3\cap\Gamma} \left(\frac{L}{N_g}\right)^3 \phi(|v_N(x)|^2) \right| \\ &\leq C_\phi \left| \sum_{x\in\frac{L}{N_g}\mathbb{Z}^3\cap\Gamma} \left(\frac{L}{N_g}\right)^3 (1+|v_N(x)|^4) \right| \\ &= C_\phi \int_\Gamma (1+|v_N|^4) = C_\phi \left(|\Gamma| + \|v_N\|^4_{L^4_\#} \right).\end{aligned}$$

This completes the proof of Lemma 2.1. □

3 Thomas-Fermi-von-Weizsäcker model

In the TFW model, as well as in any orbital-free model, the ground state electronic density of the system is obtained by minimizing an explicit functional of the density. Denoting by $\mathcal{N}$ the number of electrons in the simulation cell and by

$$\mathfrak{R}_{\mathcal{N}} = \left\{ \rho \geq 0 \mid \sqrt{\rho} \in H^1_{\#}(\Gamma),\ \int_\Gamma \rho = \mathcal{N} \right\}$$

the set of admissible densities, the TFW problem reads

$$I^{\mathrm{TFW}} = \inf \left\{ \mathcal{E}^{\mathrm{TFW}}(\rho),\ \rho \in \mathfrak{R}_{\mathcal{N}} \right\}, \tag{14}$$

where

$$\mathcal{E}^{\mathrm{TFW}}(\rho) = \frac{C_{\mathrm{W}}}{2} \int_\Gamma |\nabla \sqrt{\rho}|^2 + C_{\mathrm{TF}} \int_\Gamma \rho^{5/3} + \int_\Gamma \rho V^{\mathrm{ion}} + \frac{1}{2} D_\Gamma(\rho, \rho).$$

C_{W} is a positive real number ($C_{\mathrm{W}} = 1$, $1/5$ or $1/9$ depending on the context [**?**]), and C_{TF} is the Thomas-Fermi constant: $C_{\mathrm{TF}} = \frac{10}{3}(3\pi^2)^{2/3}$. The last term of the TFW energy models the periodic Coulomb energy: for ρ and ρ' in $H^{-1}_{\#}(\Gamma)$,

$$D_\Gamma(\rho, \rho') := 4\pi \sum_{k \in \mathcal{R}^* \setminus \{0\}} |k|^{-2} \overline{\widehat{\rho}}_k \, \widehat{\rho}'_k.$$

We finally make the assumption that V^{ion} is a periodic potential such that

$$\exists m > 3,\ C \geq 0 \text{ s.t. } \forall k \in \mathcal{R}^*,\ |\widehat{V}^{\mathrm{ion}}_k| \leq C|k|^{-m}. \tag{15}$$

Hence, we derive

$$\begin{aligned} \|V\|_{H^s_{\#}} &= \sum_{k \in \mathcal{R}^*} (1 + |k|^2)^s |\widehat{V}_k|^2 \\ &\leq \sum_{k \in \mathcal{R}^* \setminus \{0\}} (1 + |k|^2)^s \frac{1}{|k|^{2m}} \\ &\leq \sum_{k \in \mathcal{R}^* \setminus \{0\}} \frac{1}{|k|^{2(m-s)}}. \end{aligned}$$

By standard arguments of equivalence of series with integrals (over $\mathbb{R}^3$) we get the convergence of $\displaystyle\sum_{k \in \mathcal{R}^* \setminus \{0\}} \frac{1}{|k|^{2(m-s)}}$ if $2(m-s-1) > 1$, i.e. if $s < m - \frac{3}{2}$, this implies that V^{ion} is in $H^{m-3/2-\epsilon}(\Gamma)$ for all $\epsilon > 0$.

It is convenient to reformulate the TFW model in terms of $v = \sqrt{\rho}$. It can be seen that

$$I^{\mathrm{TFW}} = \inf \left\{ E^{\mathrm{TFW}}(v),\ v \in H^1_{\#}(\Gamma),\ \int_\Gamma |v|^2 = \mathcal{N} \right\} \tag{16}$$

where

$$E^{\mathrm{TFW}}(v) = \frac{C_{\mathrm{W}}}{2}\int_\Gamma |\nabla v|^2 + C_{\mathrm{TF}}\int_\Gamma |v|^{10/3} + \int_\Gamma V^{\mathrm{ion}}|v|^2 + \frac{1}{2}D_\Gamma(|v|^2,|v|^2).$$

It is well known [31] that (14) has a unique minimizer ρ^0, and that the minimizers of (16) are u and $-u$, where $u = \sqrt{\rho^0}$. Besides, the function u is in $H^{m+1/2-\epsilon}_\#(\Gamma)$ for any $\epsilon > 0$ (and therefore in $C^2_\#(\Gamma)$ since $m+1/2-\epsilon > 7/2$ for ϵ small enough), is positive everywhere in Γ and satisfies the Euler equation

$$-\frac{C_{\mathrm{W}}}{2}\Delta u + \left(\frac{5}{3}C_{\mathrm{TF}}u^{4/3} + V^{\mathrm{ion}} + V^{\mathrm{Coulomb}}_{u^2}\right)u = \lambda u$$

for some $\lambda \in \mathbb{R}$, where

$$V^{\mathrm{Coulomb}}_{\rho^0}(x) = 4\pi \sum_{k\in\mathcal{R}^*\setminus\{0\}} |k|^{-2}\widehat{\rho}^0_k e_k(x)$$

is the periodic Coulomb potential generated by the periodic charge distribution ρ. Recall that $V^{\mathrm{Coulomb}}_\rho$ can also be defined as the unique solution in $H^1_\#(\Gamma)$ to

$$\begin{cases} -\Delta V^{\mathrm{Coulomb}}_\rho = 4\pi\left(\rho - |\Gamma|^{-1}\displaystyle\int_\Gamma \rho\right) \\ \displaystyle\int_\Gamma \rho V^{\mathrm{Coulomb}}_\rho = 0. \end{cases}$$

The planewave discretization of the TFW model is obtained by choosing

1. an energy cut-off $E_{\mathrm{c}} > 0$ or, equivalently, a finite dimensional Fourier space V_{N_c}, the integer N_c being related to E_{c} through the relation $N_c := [\sqrt{2E_{\mathrm{c}}}\,L/2\pi]$;
2. a cartesian grid $\mathcal{G}_{N_g}$ with step size L/N_g where $N_g \in \mathbb{N}^*$ is such that $N_g \geq 4N_c + 1$,

and by considering the finite dimensional minimization problem

$$I^{\mathrm{TFW}}_{N_c,N_g} = \inf\left\{E^{\mathrm{TFW}}_{N_g}(v_{N_c}),\ v_{N_c}\in V_{N_c},\ \int_\Gamma |v_{N_c}|^2 = \mathcal{N}\right\}, \tag{17}$$

where

$$\begin{aligned} E^{\mathrm{TFW}}_{N_g}(v_{N_c}) \;=\; & \frac{C_{\mathrm{W}}}{2}\int_\Gamma |\nabla v_{N_c}|^2 + C_{\mathrm{TF}}\int_\Gamma \mathcal{I}_{N_g}(|v_{N_c}|^{10/3}) + \int_\Gamma \mathcal{I}_{N_g}(V^{\mathrm{ion}})|v_{N_c}|^2 \\ & + \frac{1}{2}D_\Gamma(|v_{N_c}|^2,|v_{N_c}|^2), \end{aligned}$$

$\mathcal{I}_{N_g}$ denoting the interpolation operator introduced in the previous section. The Euler equation associated with (17) can be written as a nonlinear eigenvalue problem

$$\forall v_{N_c}\in V_{N_c},\quad \langle(\widetilde{H}^{N_g}_{|u_{N_c,N_g}|^2}u_{N_c,N_g} - \lambda_{N_c,N_g})u_{N_c,N_g}, v_{N_c}\rangle_{H^{-1}_\#,H^1_\#} = 0$$

where we have denoted by

$$\widetilde{H}_\rho^{N_g} = -\frac{C_{\rm W}}{2}\Delta + \mathcal{I}_{N_g}\left(\frac{5}{3}C_{\rm TF}\rho^{2/3} + V^{\rm ion}\right) + V_\rho^{\rm Coulomb}$$

the pseudo-spectral TFW Hamiltonian associated with the density ρ, and by λ_{N_c,N_g} the Lagrange multiplier of the constraint $\int_\Gamma |v_{N_c}|^2 = \mathcal{N}$. We therefore have

$$-\frac{C_{\rm W}}{2}\Delta u_{N_c,N_g} + \Pi_{N_c}\left[\left(\mathcal{I}_{N_g}\left(\frac{5}{3}C_{\rm TF}|u_{N_c,N_g}|^{4/3} + V^{\rm ion}\right) + V^{\rm Coulomb}_{|u_{N_c,N_g}|^2}\right)u_{N_c,N_g}\right] = \lambda_{N_c,N_g}u_{N_c,N_g}.$$

Under the condition that $N_g \geq 4N_c + 1$, we have for all $\phi \in C^0_\#(\Gamma)$,

$$\forall (k,l) \in \mathcal{R}^* \times \mathcal{R}^* \text{ s.t. } |k|, |l| \leq \frac{2\pi}{L}N_c, \quad \int_\Gamma \mathcal{I}_{N_g}(\phi)\, e_k^*\, e_l = \widehat{\phi}^{\rm FFT}_{k-l},$$

so that, $\widetilde{H}_{u_{N_c,N_g}}$ is defined on V_{N_c} by the Fourier matrix

$$\begin{aligned}[\widehat{H}^{N_g}_{|u_{N_c,N_g}|^2}]_{kl} &= \frac{C_{\rm W}}{2}|k|^2\delta_{kl} + \frac{5}{3}C_{\rm TF}\widehat{(|u_{N_c,N_g}|^{4/3})}^{{\rm FFT},N_g}_{k-l} + \widehat{(V^{\rm ion})}^{{\rm FFT},N_g}_{k-l} \\ &\quad + 4\pi\frac{\widehat{(|u_{N_c,N_g}|^2)}^{{\rm FFT},N_g}_{k-l}}{|k-l|^2}(1-\delta_{kl}),\end{aligned}$$

where, by convention, the last term of the right hand side is equal to zero for $k = l$.

We also introduce the variational approximation of (16)

$$I^{\rm TFW}_{N_c} = \inf\left\{E^{\rm TFW}(v_{N_c}),\ v_{N_c} \in V_{N_c},\ \int_\Gamma |v_{N_c}|^2 = \mathcal{N}\right\}. \tag{18}$$

Any minimizer u_{N_c} to (18) satisfies the elliptic equation

$$-\frac{C_{\rm W}}{2}\Delta u_{N_c} + \Pi_{N_c}\left[\frac{5}{3}C_{\rm TF}|u_{N_c}|^{4/3}u_{N_c} + V^{\rm ion}u_{N_c} + V^{\rm Coulomb}_{|u_{N_c}|^2}u_{N_c}\right] = \lambda_{N_c}u_{N_c}, \tag{19}$$

for some $\lambda_{N_c} \in \mathbb{R}$.

Theorem 3.1 *For each $N_c \in \mathbb{N}$, we denote by u_{N_c} a minimizer to (18) such that $(u_{N_c}, u)_{L^2_\#} \geq 0$ and, for each $N_c \in \mathbb{N}$ and $N_g \geq 4N_c + 1$, we denote by u_{N_c,N_g} a minimizer to (17) such that $(u_{N_c,N_g}, u)_{L^2_\#} \geq 0$. Then for N_c large enough, u_{N_c} and u_{N_c,N_g} are unique, and the following estimates hold true, for $\epsilon > 0$*

$$\begin{aligned}
\|u_{N_c} - u\|_{H^s_\#} &\leq C_s N_c^{-(m-s+1/2-\epsilon)} && (20)\\
|\lambda_{N_c} - \lambda| &\leq C N_c^{-(2m-1-\epsilon)} && (21)\\
\gamma\|u_{N_c} - u\|^2_{H^1_\#} \leq I^{\rm TFW}_{N_c} - I^{\rm TFW} &\leq C\|u_{N_c} - u\|^2_{H^1_\#} && (22)\\
\|u_{N_c,N_g} - u_{N_c}\|_{H^s_\#} &\leq C_s\, N_c^{3/2+(s-1)_+}N_g^{-m}, && (23)\\
|\lambda_{N_c,N_g} - \lambda_{N_c}| &\leq C N_c^{3/2}N_g^{-m}, && (24)\\
|I^{\rm TFW}_{N_c,N_g} - I^{\rm TFW}_{N_c}| &\leq C N_c^{3/2}N_g^{-m}, && (25)
\end{aligned}$$

for all $-m+3/2 < s < m+1/2$ *and for some constants* $\gamma > 0$, $C \geq 0$ *and* $C_s \geq 0$ *independent of* N_c *and* N_g. *Beside* $(s-1)_+$ *denotes the positive part of* $s-1$.

Remark 3.2 *More complex orbital-free models have been proposed in the recent years [53], which are used to perform multimillion atom DFT calculations. Some of these models however are not well posed (the energy functional is not bounded below [3]), and the others are not well understood from a mathematical point of view. For these reasons, we will not deal with those models in this article.*

Proof The outline of this proof, we will be as :

1. We establish the estimates (20), (21) and (22)
2. We prove the uniqueness of u_{N_c} a minimizer to (18) such that $(u_{N_c}, u)_{L^2_\#} \geq 0$.
3. We establish the estimates (23), (24) and (25)
4. We prove the uniqueness of u_{N_c,N_g} a minimizer to (17) such that $(u_{N_c,N_g}, u)_{L^2_\#} \geq 0$.

For simplicity, we will use the following notations

- $X = H^1_\#(\Gamma)$
- $a(v,w) = \dfrac{C_W}{2}\displaystyle\int_\Gamma \nabla v \nabla w + \int_\Gamma V^{\text{ion}} vw$
- $F(t) = C_{\text{TF}}\, t^{5/3}$
- $J(u) = D_\Gamma(u^2,u^2) = 4\pi \displaystyle\sum_{k\in\mathcal{R}^*\setminus\{0\}} \frac{\widehat{\rho_k}\overline{\widehat{\rho_k}}}{|k|^2} = 4\pi \sum_{k\in\mathcal{R}^*\setminus\{0\}} \frac{|\widehat{\rho_k}|^2}{|k|^2}$
- $A_\rho w = -\dfrac{C_{\text{W}}}{2}\Delta w + F'(\rho)w + V^{\text{ion}} w + V_\rho^{\text{Coulomb}} w$

3.1 Step 1: first part of the *a priori* errors estimates

The estimates (20), (21) and (22) originate from arguments already introduced in the chapter 1. The following lemma, gathers some results which will be usefull for the proof of this estimates.

Lemma 3.3 *There exists* $C \geq 0$ *such that for all* $w \in H^1_\#(\Gamma)$,

$$0 \leq \langle (A_{\rho^0} - \lambda)w, w\rangle_{X,X'} \leq C\|w\|^2_{H^1} \tag{26}$$

Proof For all $w \in H^1_\#(\Gamma)$

$$\begin{aligned}\langle (A_{\rho^0} - \lambda)w, w\rangle_{X,X'} &= \frac{C_{\mathrm{W}}}{2}\int_\Gamma |\nabla w|^2 + \int_\Gamma F'(u^2)w^2 + \int_\Gamma V^{\mathrm{ion}} w^2 \\ &\quad + \int_\Gamma V^{\mathrm{Coulomb}}_{\rho^0} w^2 - \lambda \int_\Gamma w^2. \end{aligned} \tag{27}$$

For $m > 3$ and $\epsilon > 0$, we have that $H^{m+1/2-\epsilon}_\#(\Gamma)$ and $H^{m-3/2+\epsilon}_\#(\Gamma)$ are both algebras and embedded in $L^\infty(\Gamma)$, thereby, for all $w \in H^1_\#(\Gamma)$, we obtain the followings upper bound

$$\begin{aligned}\left|\int_\Gamma F'(u^2)w^2\right| &\leq C\|u\|^{4/3}_{L^\infty}\int_\Gamma w^2 \leq C\|w\|^2_{L^2_\#} \\ &\leq C\|w\|^2_{H^1_\#} \end{aligned} \tag{28}$$

and

$$\begin{aligned}|\int_\Gamma V^{\mathrm{ion}} w^2| &\leq C\|V^{\mathrm{ion}}\|_{L^\infty_\#}\int w^2 \leq C\|w\|^2_{L^2_\#} \\ &\leq C\|w\|^2_{H^1_\#}. \end{aligned} \tag{29}$$

In addition, by noticing that

$$\begin{aligned}\sum_{k\in\mathcal{R}^*\setminus\{0\}}\left|(\widehat{V^{\mathrm{Coulomb}}_{\rho^0}})_k\right| &\leq 4\pi\sum_{k\in\mathcal{R}^*\setminus\{0\}}\left|\frac{\widehat{(\rho^0)}_k}{|k|^2}\right| \\ &\leq \Big(\sum_{k\in\mathcal{R}^*\setminus\{0\}}\frac{1}{|k|^4}\Big)^{1/2}\Big(\sum_{k\in\mathcal{R}^*\setminus\{0\}}|\widehat{(\rho^0)}_k|^2|\Big)^{1/2} \\ &\leq C\|\rho^0\|_{L^2_\#}, \end{aligned}$$

we derive that

$$\begin{aligned}\|V^{\mathrm{Coulomb}}_{\rho^0}\|_{L^\infty_\#} &\leq \sum_{k\in\mathcal{R}^*\setminus\{0\}}\left|(\widehat{V^{\mathrm{Coulomb}}_{\rho^0}})_k\right| \\ &\leq C\|\rho^0\|_{L^2_\#}. \end{aligned} \tag{30}$$

Hence,

$$|\int_\Gamma V^{\mathrm{Coulomb}}_{\rho^0} w^2 \leq \|V^{\mathrm{Coulomb}}_{\rho^0}\|_{L^\infty_\#}\int_\Gamma w^2 \leq C\|\rho^0\|_{L^2_\#}\int_\Gamma w^2 \leq C\|\rho^0\|_{L^2_\#}\|w\|^2_{L^2_\#}$$

We obtain the following upper bound,

$$|\int_\Gamma V^{\mathrm{Coulomb}}_{\rho^0} w^2| \leq C\|\ w\|^2_{H^1_\#}. \tag{31}$$

Therefore, by combining this, with (28) and (29) in (27), we deduce that there exists a positive constante C such that

$$\forall w \in H^1_\#(\Gamma) \qquad \langle (A_{\rho^0} - \lambda)w, w\rangle_{X,X'} \leq C\|w\|^2_{H^1_\#}. \tag{32}$$

which concludes the proof of the Lemma 3.3.
□

We denote by $u^\perp$, the subspace of $H^1_\#(\Gamma)$ such that for all $w^\perp \in u^\perp$ we have $(u, w^\perp)_{L^2} = 0$. By using the fact that λ is the lowest eigenvalue of A_ρ, we get

$$\forall w^\perp \in u^\perp), \quad \langle (A_{\rho^0} - \lambda) w^\perp, w^\perp \rangle_{X,X'} \geq (\lambda_2 - \lambda_1) \|w^\perp\|^2_{L^2} \geq 0, \tag{33}$$

For all $w \in H^1_\#(\Gamma)$, we can write w as $w^\perp + u(u, w)_{L^2}$, hence

$$\begin{aligned} \langle (A_{\rho^0} - \lambda) w, w \rangle_{X,X'} &= \langle (A_{\rho^0} - \lambda) w^\perp, w^\perp \rangle_{X,X'} \\ &\geq (\lambda_2 - \lambda_1) \|w^\perp\|^2_{L^2} \geq 0 \end{aligned} \tag{34}$$

$$\langle (A_{\rho^0} - \lambda) w, w \rangle_{X,X'} \geq \frac{C_W}{2} \|\nabla w\|^2_{L^2_\#} - C(\|V^{\text{ion}}\|_{L^\infty_\#} + \|V^{\text{Coulomb}}_{\rho^0}\|_{L^\infty_\#}) \|w\|^2_{L^2_\#},$$

Hence, there exists a positive constante c such that

$$\forall w \in H^1_\#(\Gamma) \qquad \langle (A_{\rho^0} - \lambda) w, w \rangle_{X,X'} \geq \frac{C_{\text{W}}}{2} \|\nabla w\|_{L^2} - c \|w\|^2_{L^2} \tag{35}$$

Lemma 3.4 *There exists $C \geq 0$ such that for all $w \in H^1_\#(\Gamma)$,*

$$c\|w\|^2_{H^1} \leq \langle (E''(u) - 2\lambda) w, w \rangle_{X,X'} \leq C \|w\|^2_{H^1} \tag{36}$$

Proof We have for all v and $w \in H^1_\#(\Gamma)$,

$$\begin{aligned} \langle (E''(u) - 2\lambda) v, w \rangle_{X,X'} &= C_{\text{W}} \int_\Gamma \nabla w \nabla v + \frac{70}{9} C_{\text{TF}} \int_\Gamma |u|^{4/3} vw + \int_\Gamma V^{\text{ion}} vw \\ &\quad + \frac{1}{2} \langle J''(u) v, w \rangle_{X,X'} - 2\lambda \int_\Gamma vw \\ &= 2 \langle (A_{\rho^0} - \lambda) v, w \rangle_{X,X'} + 4 \int_\Gamma F''(u^2) u^2 vw + 4\, D_\Gamma(uv, uw) \end{aligned} \tag{37}$$

and

$$\langle J''(u) v, w \rangle_{X,X'} = 4\, D_\Gamma(u^2, vw) + 8\, D_\Gamma(uv, uw) \tag{38}$$

$$\tag{39}$$

Since J is positive and quadratic, it is convex, and for all $w \in H^1_\#(\Gamma)$,

$$\langle J''(u) w, w \rangle_{X,X'} \geq 0. \tag{40}$$

Besides, for all v, $w \in H^1_\#(\Gamma)$, we have

$$\begin{aligned} D_\Gamma(u^2, vw) &= \int_\Gamma V^{\text{Coulomb}}_{\rho^0} vw \leq C\, \|V^{\text{Coulomb}}_{\rho^0}\|_{L^\infty_\#} \int_\Gamma vw \\ &\leq C \|\rho^0\|_{L^2_\#} \|v\|_{L^2_\#} \|w\|_{L^2_\#} \end{aligned} \tag{41}$$

and

$$\begin{aligned} D_\Gamma(vw,vw) &= 4\pi \sum_{k\in\mathcal{R}^*\setminus\{0\}} \frac{\widehat{(vw)}_k\overline{\widehat{(vw)}_k}}{|k|^2} \\ &= 4\pi \sum_{k\in\mathcal{R}^*\setminus\{0\}} \frac{|\widehat{(vw)}_k|^2}{|k|^2} \geq 0 \qquad (42) \\ &\leq C \sum_{k\in\mathcal{R}^*\setminus\{0\}} |\widehat{(vw)}_k|^2 = C\|vw\|^2_{L^2_\#} \\ &\leq C\|v\|^2_{L^4_\#}\|w\|^2_{L^4_\#} \\ &\leq C\|v\|^2_{H^1_\#}\|w\|^2_{H^1_\#} \qquad (43) \end{aligned}$$

Therefore, by combining this with (41) in (38), we get

$$\langle J''(u)w,w\rangle_{X,X'} \leq C\|w\|^2_{H^1_\#}. \qquad (44)$$

Besides, by using (34) and (42), we obtain

$$\begin{aligned} |\langle (E''(u)-2\lambda)w,w\rangle_{X,X'}| &= 2\langle (A_{\rho^0}-\lambda)w,w\rangle_{X,X'} + 4\int_\Gamma F''(u^2)u^2ww + 4\,D_\Gamma(uw,uw) \\ &\leq C\|w\|^2_{H^1_\#} + C\|u\|^{4/3}_{L^\infty_\#}\|w\|^2_{L^2_\#} \\ &\leq C\|w\|^2_{H^1}. \qquad (45) \end{aligned}$$

To prove the lower bound of (36), we first need to establish that

$$\langle (E''(u)-2\lambda)w,w\rangle_{X,X'} \geq 0 \qquad \forall w\in H^1_\#(\Gamma). \qquad (46)$$

First we notice that $\langle (E''(u)-2\lambda)w,w\rangle_{X,X'}$ can be rewritten as

$$\begin{aligned} \langle (E''(u)-2\lambda)w,w\rangle_{X,X'} &= 2\langle (A_{\rho^0}-\lambda)w,w\rangle_{X,X'} + \frac{40}{9}C_{\mathrm{TF}}\int_\Gamma |u|^{4/3}w^2 \\ &\quad +4\,D(uw,uw) \qquad (47) \end{aligned}$$

in such way that we can use (34) and (42) to get (46).

Then by reasonning by contraction, we prove the lower bound of (36),

$$\langle (E''(u)-2\lambda)w,w\rangle_{X,X'} \geq c\|w\|^2_{H^1} \qquad \forall w\in H^1_\#(\Gamma). \qquad (48)$$

Assume (48) does not hold and thus we may get a sequence $(w_n)_{n\geq 1}$ bounded in $H^1_\#(\Gamma)$, such that$\|w_n\|^2_{H^1}=1$. Hence we can extract a sub sequence denoted by $(w_{n_k})_{n_k\geq 1}$ such that

$$\lim_{n_k\to+\infty}\langle (E''(u)-2\lambda)w_{n_k},w_{n_k}\rangle_{X,X'} = 0. \qquad (49)$$

Since $(w_n)_{n\geq 1}$ is bounded in $H^1_\#(\Gamma)$, we can extract a sub sequence weakly convergent in $H^1_\#(\Gamma)$ to w dans $H^1_\#(\Gamma)$.

Then by using that $H^1_{\#}(\Gamma)$ is embedded in all $L^p_{\#}(\Gamma)$ for $1 \leq p < 6$, we can extract a sub sequence denoted by $(w_{n_{k'}})_{n_{k'} \geq 1}$ strongly convergent in $L^p_{\#}(\Gamma)$ to w'. Hence, from the uniqueness of the limit, we get

$$w_{n_k} \underset{H^1}{\rightharpoonup} w \tag{50}$$

$$w_{n_k} \underset{L^p}{\rightarrow} w \qquad \text{for } 1 \leq p < 6, \tag{51}$$

and

$$w^2_{n_k} \underset{L^2}{\rightarrow} w^2. \tag{52}$$

Therefore

$$\lim_{n_k \to +\infty} |2\lambda \int_\Gamma [w^2 - w^2_{n_k}]| = 0, \tag{53}$$

$$\lim_{n_k \to +\infty} | \int_\Gamma |u|^{4/3}(w^2 - w^2_{n_k})| = 0, \tag{54}$$

$$\lim_{n_k \to +\infty} | \int_\Gamma V^{\text{ion}}[w^2 - w^2_{n_k}]| = 0, \tag{55}$$

Indeed,

$$\lim_{n_k \to +\infty} | \int_\Gamma |u|^{4/3}(w^2 - w^2_{n_k})| \leq \lim_{n_k \to +\infty} \|u\|^{4/3}_{L^\infty_{\#}} \|w^2 - w^2_{n_k}\|_{L^2} = 0 \quad \text{ from (52)}$$

and,

$$\lim_{n_k \to +\infty} | \int_\Gamma V^{\text{ion}}[w^2 - w^2_{n_k}]| \leq \lim_{n_k \to +\infty} \|V^{\text{ion}}\|^{4/3}_{L^\infty_{\#}} \|w^2 - w^2_{n_k}\|_{L^2} = 0 \quad \text{ from (52)}$$

By the same way, we prove that

$$\lim_{n_k \to +\infty} |\langle (J''(u)(w - w_{n_k}), w - w_{n_k} \rangle_{X,X'}| = 0. \tag{56}$$

Indeed, from (41) and (42), we derive

$$\begin{aligned} \lim_{n_k \to +\infty} |\langle (J''(u)(w - w_{n_k}), w - w_{n_k} \rangle_{X,X'}| &\leq C \lim_{n_k \to +\infty} \|u\|^2_{L^4} \|w^2 - w^2_{n_k}\|_{L^4} \\ &= 0 \qquad \text{from(51) and (52) with } p = 4. \end{aligned}$$

Since $w \mapsto C_{\text{W}} \displaystyle\int_\Gamma |\nabla w|^2$ is continuous in $H^1_{\#}(\Gamma)$ and convex, it is also weakly lower semicontinuous. Hence, using this in (37)

$$0 \leq \langle (E''(u) - 2\lambda)w, w \rangle_{X,X'} \leq \lim_{n \to +\infty} \langle (E''(u) - 2\lambda)w_n, w_n \rangle_{X,X'} = 0 \tag{57}$$

and $w = 0$.
Thereby, using this with (53) - (55) and the fact that $w \mapsto C_{\text{W}} \displaystyle\int_\Gamma |\nabla w|^2$ is weakly lower semicontinuous, we get

$$\begin{aligned} 0 = \lim_{n \to +\infty} \langle (E''(u) - 2\lambda)w_n, w_n \rangle_{X,X'} &\geq C_{\text{W}} \lim_{n \to +\infty} \inf \int_\Gamma |\nabla w_n|^2 \\ &+ \frac{70}{9} C_{\text{TF}} \lim_{n \to +\infty} \inf \int_\Gamma |u|^{4/3} w_n^2 + 0, \end{aligned}$$

Then by using (54) and the fact that $w = 0$, we derive

$$\begin{aligned} 0 = \lim_{n\to+\infty} \langle (E''(u) - 2\lambda) w_n, w_n \rangle_{X,X'} &\geq C_{\mathrm{W}} \lim_{n\to+\infty} \inf \int_\Gamma |\nabla w_n|^2 \\ &\geq C_{\mathrm{W}} \lim_{n\to+\infty} \inf \sum_{\ell\in\mathcal{R}^*} |\ell|^2 \widehat{(w_n)}_\ell \overline{\widehat{(w_n)}_\ell} \\ &\geq \frac{C_{\mathrm{W}}}{2} \lim_{n\to+\infty} \inf \sum_{\ell\in\mathcal{R}^*} (1+|\ell|^2) \widehat{(w_n)}_\ell \overline{\widehat{(w_n)}_\ell} \\ &\geq \frac{C_{\mathrm{W}}}{2} \lim_{n\to+\infty} \inf \|w_m\|^2_{H^1_\#} = 1. \end{aligned}$$

which is impossible. □

We denote by $u_{N_c} \in V_{N_c}$ a miminizer of (18) such that $(u_{N_c}, u)_{L^2} \geq 0$ and which satisfies

$$\forall\, w_{N_c} \in V_{Nc} \qquad \langle (A_{\rho_{N_c}} - \lambda_{N_c}) u_{N_c}, w_{N_c} \rangle_{X',X} = 0 \quad \text{with } \rho_{N_c} = u^2_{N_c} \tag{58}$$

for some $\lambda_{N_c} \in \mathbb{R}$. Obviously $-u_{N_c}$ is also a minimizer associated with the same eigenvalue λ_{N_c}, we will prove that those are the only ones and sastify the estimations (20)-(22).

To start we need to establish the following lower bound.

$$\gamma \|u_{N_c} - u\|^2_{H^1} \leq \langle (A_{\rho^0} - \lambda)(u_{N_c} - u), (u_{N_c} - u) \rangle_{X,X'}. \tag{59}$$

From (34), we get

$$\begin{aligned} \langle (A_{\rho^0} - \lambda) w, w \rangle_{X,X'} &\geq (\lambda_2 - \lambda_1) \left[\|w\|^2_{L^2_\#} - \frac{1}{\mathcal{N}} |(u,w)_{L^2_\#}|^2 \right] \\ &\geq (\lambda_2 - \lambda_1) \left[\|w\|^2_{L^2_\#} - |(u,w)_{L^2_\#}| \right]. \end{aligned}$$

By noticing that $\displaystyle\int_\Gamma u^2_{N_c} = \int_\Gamma u^2 = \mathcal{N}$, we get

$$\|u_{N_c}\|^2_{L^2} - |(u, u_{N_c})_{L^2}| \geq \mathcal{N} - \left| (u, u_{N_c})_{L^2_\#} \right| = \frac{1}{2} \|u - u_{N_c}\|^2_{L^2_\#}.$$

Hence,

$$\langle (A_{\rho^0} - \lambda)(u_{N_c} - u), (u_{N_c} - u) \rangle_{X,X'} \geq \frac{\lambda_2 - \lambda_1}{2} \|u - u_{N_c}\|^2_{L^2_\#}.$$

Let be θ such that $0 < \theta \leq \dfrac{\lambda_2 - \lambda_1}{\lambda_2 + \lambda_1 + 2\|V^{\rm ion}\|_{L^\infty_\#} + 2\|V^{\rm Coulomb}_{\rho^0}\|_{L^\infty_\#}} < 1$, from the previous inequality and (35), we get

$$\begin{aligned}
\langle (A_{\rho^0} - \lambda)(u_{N_c} - u), (u_{N_c} - u)\rangle_{X,X'} &\geq \theta \frac{C_W}{2}\|\nabla(u_{N_c} - u)\|^2_{L^2_\#} \\
&\quad -\theta\left(\lambda_1 + \|V^{\rm ion}\|_{L^\infty_\#} + \|V^{\rm Coulomb}_{\rho^0}\|_{L^\infty_\#}\right)\|u_{N_c} - u\|^2_{L^2_\#} \\
&\quad +(1-\theta)\frac{\lambda_2 - \lambda_1}{2}\|u_{N_c} - u\|^2_{L^2_\#} \\
&\geq \theta \frac{C_W}{2}\|\nabla(u_{N_c} - u)\|^2_{L^2_\#} + \frac{\lambda_2 - \lambda_1}{2}\|u_{N_c} - u\|^2_{L^2_\#} \\
&\quad -\frac{\theta}{2}\left(\lambda_2 + \lambda_1 + 2\|V^{\rm ion}\|_{L^\infty_\#} + 2\|V^{\rm Coulomb}_{\rho^0}\|_{L^\infty_\#}\right)\|u_{N_c} - u\|^2_{L^2_\#} \\
&\geq C\|u_{N_c} - u\|^2_{H^1_\#}.
\end{aligned}$$

In addition, we have the following convergence result

Lemma 3.5 *Let be $u_{N_c} \in V_{N_c}$ a minimizer of (18) such that $(u, u_{N_c})_{L^2_\#} \geq 0$ and that verifies (58) for some $\lambda_{N_c} \in \mathbb{R}$ then,*

$$\|u - u_{N_c}\|_{H^1_\#} \longrightarrow 0. \tag{60}$$

Proof To obtain this, we first use the fact that

$$\begin{aligned}
E(u_{N_c}) - E(u) &= \langle A_{\rho^0} u_{N_c}, u_{N_c}\rangle_{X',X} - \langle A_{\rho^0} u, u\rangle_{X',X} \\
&\quad + \int_\Gamma \left[F(u^2_{N_c}) - F(u^2) - F'(u^2)(u^2_{N_c} - u^2)\right] \\
&\quad + 2\pi \sum_{k\in\mathcal{R}^*\setminus\{0\}} \frac{\left[\widehat{(u^2_{N_c})_k}\,\overline{\widehat{(u^2_{N_c})}}_k - \widehat{(u^2)_k}\,\overline{\widehat{(u^2)}}_k - 2\widehat{(u^2)_k}\left\{\overline{\widehat{(u^2_{N_c})}} - \overline{\widehat{(u^2)}}_k\right\}\right]}{|k|^2} \\
&= \langle (A_{\rho^0} - \lambda)(u_{N_c} - u), u_{N_c} - u\rangle_{X',X} \\
&\quad + \int_\Gamma \left[F(u^2_{N_c}) - F(u^2) - F'(u^2)(u^2_{N_c} - u^2)\right] \\
&\quad + 2\pi \sum_{k\in\mathcal{R}^*\setminus\{0\}} \frac{|\widehat{(u^2_{N_c} - u^2)_k}|^2}{|k|^2}.
\end{aligned} \tag{61}$$

Then, combining it with the convexity of F and (59), we get

$$I_{u_{N_c}} - I = E(u_{N_c}) - E(u) \geq \frac{\eta}{2}\|u - u_{N_c}\|^2_{H^1_\#}.$$

Let $\Pi_{N_c} u \in V_{N_c}$ be such that

$$\|u - \Pi_{N_c} u\|_{H^1} = \min\{\, \|u - w_{N_c}\|_{H^1_\#},\, w_{N_c} \in V_{N_c}\} \underset{N_c\to+\infty}{\longrightarrow} 0.$$

We defined by $\tilde{u}_{N_c} = \mathcal{N}^{1/2}\Pi_{N_c}u/\|\Pi_{N_c}u\|_{L^2_\#}$, in such way that $\|\tilde{u}_{N_c}\|_{L^2_\#} = \sqrt{\mathcal{N}}$, therefore we also have

$$\|u - \tilde{u}_{N_c}\|_{H^1_\#} \underset{N_c\to+\infty}{\longrightarrow} 0.$$

The functional E being strongly continuous on $H^1_\#(\Gamma)$, we get

$$\|u - u_{N_c}\|^2_{H^1_\#} \leq \frac{2}{\eta}(E(u_{N_c}) - E(u)) \leq \frac{2}{\eta}(E(\tilde{u}_{N_c}) - E(u)) \underset{N_c\to+\infty}{\longrightarrow} 0.$$

Hence, we finally obtain (60) and we also have

$$\begin{aligned}
\lambda_{N_c} &= \mathcal{N}^{-1}\left[\frac{1}{2}\int_\Gamma |\nabla u_{N_c}|^2 + \int_\Gamma f(|u_{N_c}|^2)|u_{N_c}|^2 + \int_\Gamma V^{\rm ion}|u_{N_c}|^2 + D_\Gamma(|u_{N_c}|^2, |u_{N_c}|^2)\right] \\
&\underset{N_c\to\infty}{\longrightarrow} \mathcal{N}^{-1}\left[\frac{1}{2}\int_\Gamma |\nabla u|^2 + \int_\Gamma f(|u|^2)|u|^2 + \int_\Gamma V^{\rm ion}|u|^2 + D_\Gamma(|u|^2, |u|^2)\right] \\
&= \lambda.
\end{aligned}$$

□

We now observes that $u_{N_c}$is solution to the elliptic equation

$$-\frac{C_W}{2}\Delta u_{N_c} + \Pi_{N_c}\left[\left(V^{\rm ion} + \frac{5}{3}C_{\rm TF}u_{N_c}^{4/3} + V^{\rm Coulomb}_{\rho_{N_c}}\right)u_{N_c}\right] = \lambda_{N_c}u_{N_c} \qquad (62)$$

As $(V^{\rm ion} + \frac{5}{3}C_{\rm TF}u_{N_c}^{4/3} + V^{\rm Coulomb}_{|u_{N_c}|^2})u_{N_c}$ is bounded in $L^2_\#(\Gamma)$, uniformly in N_c, we deduce from (62) that the sequence $(u_{N_c})_{N_c\in\mathbb{N}}$ is bounded in $H^2_\#(\Gamma)$, hence in $L^\infty(\Gamma)$. In addition,

$$\begin{aligned}
\Delta\,(u_{N_c} - u) &= \Pi_{N_c}\left(V^{\rm ion}(u_{N_c} - u) + F'(u_{N_c}^2)u_{N_c} - F'(u^2)u + V^{\rm Coulomb}_{\rho_{N_c}}u_{N_c} - V^{\rm Coulomb}_{\rho}u\right) \\
&\quad -(I - \Pi_{N_c})(Vu + F'(u^2)u + V^{\rm Coulomb}_{\rho}u) - \lambda_{N_c}(u_{N_c} - u) - (\lambda_{N_c} - \lambda)u.
\end{aligned} \qquad (63)$$

As $(u_{N_c})_{N_c\in\mathbb{N}}$ is in bounded in $L^\infty(\Gamma)$ and converges to u in $H^1_\#(\Gamma)$, the right hand side of the above equality converges to 0 in $L^2_\#(\Gamma)$, which implies that $(u_{N_c})_{N_c\in\mathbb{N}}$ converges to u in $H^2_\#(\Gamma)$, and therefore in $C^0_\#(\Gamma)$.
Since $(V^{\rm ion} + \frac{5}{3}C_{\rm TF}u_N^{4/3} + V^{\rm Coulomb}_{\rho_{N_c}})$ is in $L^\infty(\Gamma)$ and that $V^{\rm ion} \in H^{m-3/2-\epsilon}$ we get that u in $H^{m+1/2-\epsilon}_\#(\Gamma)$, for all $\epsilon > 0$, by using a boostrap argument in (62).

The upper bound in (22) is obtained from (61), remarking that

$$\begin{aligned}
0 &\leq \int_\Gamma F(|u_{N_c}|^2) - F(|u|^2) - f(|u|^2)(|u_{N_c}|^2 - |u|^2) \\
&\leq \frac{35}{9}C_{\rm TF}\int_\Gamma \max(|u_{N_c}|^{4/3}, |u|^{4/3})|u_{N_c} - u|^2 \\
&\leq \frac{35}{9}C_{\rm TF}\left(\max_{N_c\in\mathbb{N}}\|u_{N_c}\|_{L^\infty}\right)^{4/3}\|u_{N_c} - u\|^2_{L^2_\#}
\end{aligned}$$

and that

$$\begin{aligned} 0 &\leq D_\Gamma(|u_{N_c}|^2 - |u|^2, |u_{N_c}|^2 - |u|^2) \leq C\||u_{N_c}|^2 - |u|^2\|_{L^2_\#}^2 \\ &\leq 4C \left(\max_{N_c \in \mathbb{N}} \|u_{N_c}\|_{L^\infty}\right)^2 \|u_{N_c} - u\|_{L^2_\#}^2 . \end{aligned}$$

Let us now establish the rates of convergence of $|\lambda_{N_c} - \lambda|$ and $\|u_{N_c} - u\|_{H^s_\#}$. As in the Chapter 1, we denote by $\psi_{u_{N_c}-u}$ the unique solution to the adjoint problem

$$\left\{ \begin{array}{l} \text{find } \psi_{u_{N_c}-u} \in u^\perp \text{ such that} \\ \forall v \in u^\perp, \quad \langle (E''(u) - 2\lambda)\psi_{u_{N_c}-u}, v\rangle_{X',X} = \langle u_{N_c} - u, v\rangle_{X',X}, \end{array} \right. \tag{64}$$

where

$$u^\perp = \left\{ v \in X \mid \int_\Gamma uv = 0 \right\}.$$

The existence and uniqueness of the solution to (64) is a straightforward consequence of (45), (48) and the Lax-Milgram lemma. Besides, $\psi_{u_{N_c}-u} \in H^2_\#(\Gamma)$ and sastifies

$$\|\psi_{u_{N_c}-u}\|_{H^1} \leq \beta^{-1} M \|u_{N_c} - u\|_{X'} \leq \beta^{-1} M \|u_{N_c} - u\|_{L^2}, \tag{65}$$

and $\|\psi_{u_{N_c}-u}\|_{H^2} \leq C\|u_{N_c} - u\|_{L^2}$. Hence,

$$\|\psi_{u_{N_c}-u} - \Pi_{N_c}\psi_{u_{N_c}-u}\|_{H^1} \leq \frac{1}{N_c}\|\psi_{u_{N_c}-u}\|_{H^2} \leq \frac{C}{N_c}\|u_{N_c} - u\|_{L^2}. \tag{66}$$

Lemma 3.6 *There exists $C \in \mathbb{R}_+$ such that for all $N_c > 0$,*

$$E(u_{N_c}) - E(u) \leq \frac{M}{2}\|u_{N_c} - u\|_{H^1}^2, \tag{67}$$

and for all $0 \leq r < m - 3/2$,

$$|\lambda_{N_c} - \lambda| \leq C\left(\|u_{N_c} - u\|_{H^1}^2 + \|u_{N_c} - u\|_{H^{-r}}\right). \tag{68}$$

Besides, there exists $N_0 > 0$ and $C \in \mathbb{R}_+$ such that for all $0 < N_c < N_0$,

$$\|u_{N_c} - u\|_{H^1} \leq C \min_{v_{N_c} \in V_{N_c}} \|v_{N_c} - u\|_{H^1} \tag{69}$$

$$\|u_{N_c} - u\|_{L^2} \leq C\left(\|u_{N_c} - u\|_{H^1}^2 + N_c^{-1}\|u_{N_c} - u\|_{H^1}\right). \tag{70}$$

Proof The upper bound in (67) is obtained from (61), remarking that

$$\begin{aligned} 0 \leq \int_\Gamma F(|u_{N_c}|^2) - F(|u|^2) - F'(|u|^2)(|u_{N_c}|^2 - |u|^2) \\ &= C_{\mathrm{TF}} \int_\Gamma |u_{N_c}|^{10/3} - |u|^{10/3} - \frac{5}{3}|u|^{4/3}(|u_{N_c}|^2 - |u|^2) \\ &\leq \frac{35}{9} C_{\mathrm{TF}} \int_\Gamma \max(|u_{N_c}|^{4/3}, |u|^{4/3})|u_{N_c} - u|^2 \\ &\leq C\|u_{N_c} - u\|_{L^2}^2 \end{aligned}$$

then that

$$0 \leq \sum_{k\in\mathcal{R}^*\setminus\{0\}} \frac{|\widehat{(u_{N_c}^2 - u^2)}_k|^2}{|k|^2} \leq C\|u_{N_c}^2 - u^2\|_{L^2}^2 \leq C\|u_{N_c} - u\|_{L^2}^2,$$

and finally by using (26).

Next, we remark that

$$\begin{aligned}
\mathcal{N}(\lambda_{N_c} - \lambda) &= \langle E'(u_{N_c}), u_{N_c}\rangle_{X',X} - \langle E'(u), u\rangle_{X',X} \\
&= a(u_{N_c}, u_{N_c}) - a(u,u) + \int_\Gamma F'(u_{N_c}^2)u_{N_c}^2 - \int_\Gamma F'(u^2)u^2 \\
&\quad + D(u_{N_c}^2, u_{N_c}^2) - D(u^2,u^2) \\
&= a(u_{N_c} - u, u_{N_c} - u) + 2a(u, u_{N_c} - u) + \int_\Gamma F'(u_{N_c}^2)u_{N_c}^2 - \int_\Gamma F'(u^2)u^2 \\
&\quad + D(u_{N_c}^2, u_{N_c}^2) - D(u^2,u^2) \\
&= a(u_{N_c} - u, u_{N_c} - u) + 2\lambda\int_\Gamma u(u_{N_c} - u) - 2\int_\Gamma F'(u^2)u(u_{N_c} - u) \\
&\quad + \int_\Gamma F'(u_{N_c}^2)u_{N_c}^2 - \int_\Gamma F'(u^2)u^2 + D(u_{N_c}^2, u_{N_c}^2) - D(u^2,u^2) \\
&\quad - 2D(u^2, u(u_{N_c} - u)) \\
&= a(u_{N_c} - u, u_{N_c} - u) - \lambda\|u_{N_c} - u\|_{L^2}^2 - 2\int_\Gamma F'(u^2)u(u_\delta - u) \\
&\quad + \int_\Gamma F'(u_{N_c}^2)u_{N_c}^2 - \int_\Gamma F(u^2)u^2 + D(u_{N_c}^2, u_{N_c}^2) - D(u^2,u^2) \\
&\quad - 2D(u^2, u(u_{N_c} - u)) \\
&= \langle (A_u - \lambda)(u_{N_c} - u), (u_{N_c} - u)\rangle_{X',X} + \int_\Gamma w_{u,u_\delta}^F (u_{N_c} - u) \\
&\quad + D((u_{N_c}^2 - u^2), u_{N_c}^2) \qquad (71)
\end{aligned}$$

where

$$w_{u,u_{N_c}}^F = u_{N_c}^2 \frac{F'(u_{N_c}^2) - F'(u^2)}{u_{N_c} - u}.$$

We know that the sequence $(u_{N_c})_{N_c\in\mathbb{N}}$ converges to u in $H_\#^{m+1/2-\epsilon}(\Gamma)$, for all $\epsilon > 0$, and that $u > 0$ in $\mathbb{R}^3$. Consequently, for N_c large enough, the function u_{N_c} (which is continuous and $\mathcal{R}$-periodic) is bounded away from 0, uniformly in N_c. As $F' \in C^\infty((0,+\infty))$, the function u_{N_c} is uniformly bounded in $H_\#^{m-3/2-\epsilon}(\Gamma)$, for all $\epsilon > 0$, (at least for N_c large enough). Hence, for all $0 \leq r < m - 3/2$

$$|\int_\Gamma w_{u,u_\delta}^F (u_{N_c} - u)| \leq C\|u - u_{N_c}\|_{H^{-r}}. \qquad (72)$$

We remark that

$$\begin{aligned} D((u_{N_c}^2-u^2),u_{N_c}^2) &= 4\pi \sum_{k,\ell\in\mathcal{R}^*\setminus\{0\}} \frac{\overline{\widehat{(u_{N_c}-u)}_\ell}\,\overline{\widehat{(u_{N_c}+u)}}_{k-\ell}\widehat{(u_{N_c}^2)}_k}{|\ell|^2} \\ &= 4\pi \sum_{k,\ell\in\mathcal{R}^*\setminus\{0\}} \frac{\overline{\widehat{(u_{N_c}-u)}_\ell}\,\widehat{(u_{N_c}+u)}_{\ell-k}\widehat{(u_{N_c}^2)}_k}{|\ell|^2} \\ &= \int_\Gamma (u_{N_c}-u)\left(V_\rho^{\mathrm{Coulomb}}(u_{N_c}+u)u_{N_c}^2\right) \end{aligned}$$

and denote by $w^J_{u,u_{N_c}} = V_\rho^{\mathrm{Coulomb}}(u_{N_c}+u)$. Since $V_\rho^{\mathrm{Coulomb}} \in L^\infty_\#(\Gamma)$ and the function u_{N_c} is uniformly bounded in $H^{m-3/2-\epsilon}_\#(\Gamma)$, for all $\epsilon>0$, (at least for N_c large enough), for all $0\le r<m-3/2$, we derive

$$\int_\Gamma (u_{N_c}-u)\left(V_\rho^{\mathrm{Coulomb}}(u_{N_c}+u)u_{N_c}^2\right)| \le C\|u-u_{N_c}\|_{H^{-r}}.$$

Then by combining this with (26) and (72) in (71), we finally get (68).

In order to evaluate the H^1-norm of the error $u_{N_c}-u$, we first notice that

$$\forall v_{N_c}\in V_{N_c},\quad \|u_{N_c}-u\|_{H^1} \le \|u_{N_c}-v_{N_c}\|_{H^1} + \|v_{N_c}-u\|_{H^1}, \tag{73}$$

and that

$$\begin{aligned} \|u_{N_c}-v_{N_c}\|_{H^1}^2 &\le \frac{1}{\beta}\langle (E''(u)-2\lambda)(u_{N_c}-v_{N_c}),(u_{N_c}-v_{N_c})\rangle_{X',X} \\ &= \frac{1}{\beta}\Big(\langle (E''(u)-2\lambda)(u_{N_c}-u),(u_{N_c}-v_{N_c})\rangle_{X',X} \\ &\qquad +\langle (E''(u)-2\lambda)(u-v_{N_c}),(u_{N_c}-v_{N_c})\rangle_{X',X}\Big). \end{aligned} \tag{74}$$

We will begin by evaluating the first term of the right hand side of (74). For all $w_{N_c}\in V_{N_c}$,

$$\begin{aligned} &\langle (E''(u)-2\lambda)(u_{N_c}-u),w_{N_c}\rangle_{X',X} \\ &= -\int_\Gamma \left(F'(u_{N_c}^2)u_{N_c}-F'(u^2)u_{N_c}-2F''(u^2)u^2(u_{N_c}-u)\right)w_{N_c} \\ &\quad + \sum_{k,\ell\in\mathcal{R}^*\setminus\{0\}} \frac{\left[\widehat{[(u^2)}_k - \widehat{(u_{N_c}^2)}_k]\overline{\widehat{(u_{N_c})}}_{k-\ell} - 2[u\widehat{(u_{N_C}-u)}]_k\overline{\widehat{(u)}}_{k-\ell}\right]\overline{\widehat{(w_{N_c})}}_\ell}{|k|^2} \\ &\quad + (\lambda_{N_c}-\lambda)\int_\Gamma u_{N_c}w_{N_c}. \end{aligned} \tag{75}$$

Besides,

$$\left|\sum_{k,\ell\in\mathcal{R}^*\setminus\{0\}} \frac{\left[[\widehat{(u^2)}_k - \widehat{(u^2_{N_c})}_k]\overline{\widehat{(u_{N_c})}}_{k-\ell} - 2(u\widehat{(u_{N_C}-u)}_k\overline{\widehat{(u)}}_{k-\ell}\right]\overline{\widehat{(w_{N_c})}}_\ell}{|k|^2}\right|$$

$$\leq \sum_{k,\ell\in\mathcal{R}^*\setminus\{0\}}\left|\frac{[[\widehat{u_{N_c}-u})^2]_k\overline{\widehat{(u_{N_c})}}_{k-\ell}\overline{\widehat{(w_{N_c})}}_\ell}{|k|^2}\right| + 2\sum_{k,\ell\in\mathcal{R}^*\setminus\{0\}}\left|\frac{[u\widehat{(u_{N_C}-u)}]_k\overline{\widehat{(u-u_{N_c})}}_{k-\ell}\overline{\widehat{(w_{N_c})}}_\ell}{|k|^2}\right|$$

$$\leq C\Big((u_{N_c}-u)^2, u_{N_c}w_{N_c}\Big)_{L^2} + 2\Big(u(u_{N_c}-u), (u_{N_c}-u)w_{N_c}\Big)_{L^2_\#}$$

$$\leq C\|u-u_{N_C}\|^2_{L^4_\#}\|w_{N_c}\|_{L^4_\#}$$

$$\leq C\|u-u_{N_C}\|^2_{H^1_\#}\|w_{N_c}\|_{H^1_\#}. \tag{76}$$

In addition, we have

$$\begin{aligned}\left|F'(u^2_{N_c})u_{N_c} - F'(u^2)u_{N_c} - 2F''(u^2)u^2(u_{N_c}-u)\right| &= \frac{5}{3}C_{\mathrm{TF}}\left|u^{7/3}_{N_c} - u^{4/3}u_{N_c} - \frac{4}{3}u^{4/3}(u_{N_c}-u)\right| \\ &\leq \frac{70}{9}C_{\mathrm{TF}}\max(|u_{N_c}|^{1/3}, |u|^{1/3})\,|u_{N_c}-u|^2\,.\end{aligned} \tag{77}$$

On the other hand, we have for all $v_{N_c}\in V_{N_c}$ such that $\|v_{N_c}\|_{L^2}=\sqrt{\mathcal{N}}$,

$$\int_\Gamma u_{N_c}(u_{N_c}-v_{N_c}) = \mathcal{N} - \int_\Gamma u_{N_c}v_{N_c} = \frac{1}{2}\|u_{N_c}-v_{N_c}\|^2_{L^2}.$$

Hence by combining this with (68) in (75) we obtain, that for all $0<N_c\leq N_1$ and all $v_{N_c}\in V_{N_c}$ such that $\|v_{N_c}\|_{L^2}=\sqrt{\mathcal{N}}$,

$$\begin{aligned}\left|\langle(E''(u)-2\lambda)(u_{N_c}-u),(u_{N_c}-v_{N_c})\rangle_{X',X}\right| \leq\ & C\Big(\|u-u_{N_c}\|^2_{H^1}\|u_{N_c}-v_{N_c}\|_{H^1} \\ & + \left(\|u_{N_c}-u\|^2_{H^1} + \|u_{N_c}-u\|_{L^2}\right)\|u_{N_c}-v_{N_c}\|^2_{L^2}\Big).\end{aligned} \tag{78}$$

It then follows from (45), (74) and (78) that for all $0<N_c\leq N_1$ and all $v_{N_c}\in X_{N_c}$ such that $\|v_{N_c}\|_{L^2}=\sqrt{\mathcal{N}}$,

$$\|u_{N_c}-v_{N_c}\|_{H^1} \leq C\left(\|u_{N_c}-u\|^2_{H^1} + \|u_{N_c}-u\|_{H^1}\|u_{N_c}-v_{N_c}\|_{H^1} + \|v_{N_c}-u\|_{H^1}\right).$$

By using this in (73) we obtain that there exists $0<N_2\leq N_1$ and $C\in\mathbb{R}_+$ such that for all $0<N_c\leq N_2$ and all $v_{N_c}\in V_{N_c}$ such that $\|v_{N_c}\|_{L^2}=\sqrt{\mathcal{N}}$,

$$\|u_{N_c}-u\|_{H^1}\leq C\|v_{N_c}-u\|_{H^1}.$$

Hence, for all $0<N_c\leq N_2$

$$\|u_{N_c}-u\|_{H^1}\leq CR_{N_c} \qquad \text{where} \qquad R_{N_c} = \min_{v_{N_c}\in v_{N_c}\,|\,\|v_{N_c}\|_{L^2}=\sqrt{\mathcal{N}}}\|v_{N_c}-u\|_{H^1}.$$

We now denote by

$$\widetilde{R}_{N_c} = \min_{v_{N_c} \in V_{N_c}} \|v_{N_c} - u\|_{H^1},$$

and by $u^0_{N_c}$ a minimizer of the above minimization problem. We know that $u^0_{N_c}$ converges to u in H^1 when N_c goes to infty. Besides,

$$\begin{aligned} R_{N_c} &\leq \|u^0_{N_c}/\|u^0_{N_c}\|_{L^2} - u\|_{H^1} \\ &\leq \|u^0_{N_c} - u\|_{H^1} + \frac{\|u^0_{N_c}\|_{H^1}}{\|u^0_{N_c}\|_{L^2}} \left|1 - \|u^0_{N_c}\|_{L^2}\right| \\ &\leq \|u^0_{N_c} - u\|_{H^1} + \frac{\|u^0_{N_c}\|_{H^1}}{\|u^0_{N_c}\|_{L^2}} \|u - u^0_{N_c}\|_{L^2} \\ &\leq \left(1 + \frac{\|u^0_{N_c}\|_{H^1}}{\|u^0_{N_c}\|_{L^2}}\right) \widetilde{R}_{N_c}. \end{aligned}$$

For $0 < N_c \leq N_2 \leq N_1$, we have $\|u^0_{N_c} - u\|_{H^1} \leq \|u_{N_c} - u\|_{H^1} \leq 1/2$, and therefore $\|u^0_{N_c}\|_{H^1} \leq \|u\|_{H^1} + 1/2$ and $\|u^0_{N_c}\|_{L^2} \geq 1/2$, yielding $R_{N_c} \leq 2(\|u\|_{H^1} + 1)\widetilde{R}_{N_c}$. Thus (69) is proved.

Let $u^*_{N_c}$ be the orthogonal projection, for the L^2 inner product, of u_{N_c} on the affine space $\left\{ v \in L^2(\Gamma) \,|\, \int_\Gamma uv = \mathcal{N} \right\}$. One has

$$u^*_{N_c} \in H^1_\#(\Gamma), \qquad u^*_{N_c} - u \in u^\perp, \qquad u^*_{N_c} - u_{N_c} = \frac{1}{2\mathcal{N}} \|u_{N_c} - u\|^2_{L^2} u,$$

from which we infer that

$$\begin{aligned} \|u_{N_c} - u\|^2_{L^2} &= \int_\Gamma (u_{N_c} - u)(u^*_{N_c} - u) + \int_\Gamma (u_{N_c} - u)(u_{N_c} - u^*_{N_c}) \\ &= \int_\Gamma (u_{N_c} - u)(u^*_{N_c} - u) - \frac{1}{2\mathcal{N}} \|u_{N_c} - u\|^2_{L^2} \int_\Gamma (u_{N_c} - u)u \\ &= \int_\Gamma (u_{N_c} - u)(u^*_{N_c} - u) + \frac{1}{2\mathcal{N}} \|u_{N_c} - u\|^2_{L^2} \left(\mathcal{N} - \int_\Gamma u_{N_c} u\right) \\ &= \int_\Gamma (u_{N_c} - u)(u^*_{N_c} - u) + \frac{1}{4\mathcal{N}} \|u_{N_c} - u\|^4_{L^2} \\ &= \langle u_{N_c} - u, u^*_{N_c} - u \rangle_{X',X} + \frac{1}{4\mathcal{N}} \|u_{N_c} - u\|^4_{L^2} \\ &= \langle (E''(u) - 2\lambda)\psi_{u_{N_c}-u}, u^*_{N_c} - u \rangle_{X',X} + \frac{1}{4\mathcal{N}} \|u_{N_c} - u\|^4_{L^2} \\ &= \langle (E''(u) - 2\lambda)(u_{N_c} - u), \psi_{u_{N_c}-u} \rangle_{X',X} \\ &\quad + \frac{1}{2\mathcal{N}} \|u_{N_c} - u\|^2_{L^2} \langle (E''(u) - 2\lambda)u, \psi_{u_{N_c}-u} \rangle_{X',X} + \frac{1}{4\mathcal{N}} \|u_{N_c} - u\|^4_{L^2} \\ &= \langle (E''(u) - 2\lambda)(u_{N_c} - u), \psi_{u_{N_c}-u} \rangle_{X',X} \\ &\quad + \|u_{N_c} - u\|^2_{L^2} \left[\int_\Gamma F''(u^2)u^3 \psi_{u_{N_c}-u} + 2 \sum_{k,\ell \in \mathcal{R}^* \setminus \{0\}} \frac{\widehat{(u^2)}_k \overline{\widehat{(u)}}_{k-\ell} \overline{\widehat{(\psi_{u_{N_c}-u})}}_\ell}{|k|^2} \right] \\ &\quad + \frac{1}{4\mathcal{N}} \|u_{N_c} - u\|^4_{L^2}. \end{aligned}$$

For all $\psi_{N_c} \in V_{N_c}$, it therefore holds

$$\begin{aligned}\|u_{N_c}-u\|_{L^2}^2 &= \langle (E''(u)-2\lambda)(u_{N_c}-u), \psi_{N_c}\rangle_{X',X} \\ &\quad +\langle (E''(u)-2\lambda)(u_{N_c}-u), \psi_{u_{N_c}-u}-\psi_{N_c}\rangle_{X',X} \\ &\quad +\|u_{N_c}-u\|_{L^2}^2\left[\int_\Gamma F'''(u^2)u^3\psi_{u_{N_c}-u} + 2\sum_{k,\ell\in\mathcal{R}^*\setminus\{0\}} \frac{\widehat{(u^2)}_k\overline{\widehat{(u)}}_{k-\ell}\overline{\widehat{(\psi_{u_{N_c}-u})}}_\ell}{|k|^2}\right] \\ &\quad +\frac{1}{4\mathcal{N}}\|u_{N_c}-u\|_{L^2}^4. \end{aligned} \tag{79}$$

From (75), we obtain that for all $\psi_{N_c} \in V_{N_c} \cap u^\perp$,

$$\begin{aligned}&\langle (E''(u)-2\lambda)(u_{N_c}-u), (\psi_{u_{N_c}-u}-\psi_{N_c})\rangle_{X',X} \\ &= -\int_\Gamma \left(F'(u_{N_c}^2)u_{N_c} - F'(u^2)u_{N_c} - 2F''(u^2)u^2(u_{N_c}-u)\right)(\psi_{u_{N_c}-u}-\psi_{N_c}) \\ &\quad +\sum_{k,\ell\in\mathcal{R}^*\setminus\{0\}} \frac{\left[[\widehat{[u_{N_c}-u)^2]}_k\overline{\widehat{(u_{N_c})}}_{k-\ell} - 2[\widehat{u(u_{N_C}-u)}]_k\overline{\widehat{(u-u_{N_c})}}_{k-\ell}\right]\overline{\widehat{(\psi_{u_{N_c}-u}-\psi_{N_c})}}_\ell}{|k|^2} \\ &\quad +(\lambda_{N_c}-\lambda)\int_\Gamma u_{N_c}(\psi_{u_{N_c}-u}-\psi_{N_c}) \end{aligned} \tag{80}$$

and

$$\begin{aligned}&\langle (E''(u)-2\lambda)(u_{N_c}-u), \psi_{N_c}\rangle_{X',X} \\ &= -\int_\Gamma \left(F'(u_{N_c}^2)u_{N_c} - F'(u^2)u_{N_c} - 2F''(u^2)u^2(u_{N_c}-u)\right)\psi_{N_c} \\ &\quad +\sum_{k,\ell\in\mathcal{R}^*\setminus\{0\}} \frac{\left[[\widehat{[u_{N_c}-u)^2]}_k\overline{\widehat{(u_{N_c})}}_{k-\ell} - 2[\widehat{u(u_{N_C}-u)}]_k\overline{\widehat{(u-u_{N_c})}}_{k-\ell}\right]\overline{\widehat{(\psi_{N_c})}}_\ell}{|k|^2} \\ &\quad +(\lambda_{N_c}-\lambda)\int_\Gamma (u_{N_c}-u)\psi_{N_c} \text{ (using that } \int_\Gamma u\psi_{N_c}=0). \end{aligned} \tag{81}$$

Therefore by using (68), (76) and (77) in (80) and (81) we get

$$\begin{aligned}\left|\langle (E''(u)-2\lambda)(u_{N_c}-u), \psi_{N_c}\rangle_{X',X}\right| &\le C\Big(\|u_{N_c}-u\|_{H^1}^2 \\ &\quad +\|u_{N_c}-u\|_{L^2}\left(\|u_{N_c}-u\|_{H^1}^2+\|u_{N_c}-u\|_{L^2}\right)\Big)\|\psi_{N_c}\|_{H^1} \end{aligned} \tag{82}$$

and

$$\begin{aligned}\left|\langle (E''(u)-2\lambda)(u_{N_c}-u), (\psi_{u_{N_c}-u}-\psi_{N_c})\rangle_{X',X}\right| &\le C\Big(\|u_{N_c}-u\|_{H^1}^2 + \\ &\quad +\|u_{N_c}-u\|_{L^2}\Big)\|\psi_{u_{N_c}-u}-\psi_{N_c}\|_{H^1}. \end{aligned} \tag{83}$$

Let $\psi^0_{N_c} \in V_{N_c} \cap u^\perp$ be such that

$$\|\psi_{u_{N_c}-u} - \psi^0_{N_c}\|_{H^1} = \min_{\psi_{N_c} \in X_{N_c} \cap u^\perp} \|\psi_{u_{N_c}-u} - \psi_{N_c}\|_{H^1}.$$

Hence, by using the fact that there exists a non negative constant C such that $\|\psi^0_{N_c}\|_{H^1} \leq \|\psi_{u_{N_c}-u}\|_{H^1} \leq C\|u_{N_c} - u\|_{L^2}$, we obtain from (45) and (82) that there exists $C \in \mathbb{R}_+$ such that for all $0 < N_c \leq N_1$,

$$\begin{aligned}\|u_{N_c} - u\|^2_{L^2} \;\leq\; & C\Big(\|u_{N_c} - u\|_{L^2} \\ & \times \left(\|u_{N_c} - u\|^2_{H^1} + \|u_{N_c} - u\|_{L^2}\left(\|u_{N_c} - u\|^2_{H^1} + \|u_{N_c} - u\|_{L^2}\right)\right) \\ & + \|u_{N_c} - u\|_{H^1}\|\psi_{u_{N_c}-u} - \psi^0_{N_c}\|_{H^1} + \|u_{N_c} - u\|^3_{L^2} + \|u_{N_c} - u\|^4_{L^2}\Big).\end{aligned}$$

Therefore, there exists $0 < N_0 \leq N_2$ and $C \in \mathbb{R}_+$ such that for all $0 < N_c \leq N_0$,

$$\|u_{N_c} - u\|^2_{L^2} \leq C\left(\|u_{N_c} - u\|_{L^2}\|u_{N_c} - u\|^2_{H^1} + \|u_{N_c} - u\|_{H^1}\|\psi_{u_{N_c}-u} - \psi^0_{N_c}\|_{H^1}\right).$$

Lastly, denoting by $\Pi^0_{V_{N_c}}$ the orthogonal projector on V_{N_c} for the L^2 inner product, a simple calculation leads to

$$\forall v \in u^\perp, \quad \min_{v_{N_c} \in V_{N_c} \cap u^\perp} \|v_{N_c} - v\|_{H^1} \leq \left(1 + \frac{\|\Pi^0_{V_{N_c}} u\|_{H^1}}{\|\Pi^0_{V_{N_c}} u\|^2_{L^2}}\right) \min_{v_{N_c} \in V_{N_c}} \|v_{N_c} - v\|_{H^1}. \tag{84}$$

Therefore, by using (66), we get

$$\|u_{N_c} - u\|^2_{L^2} \leq C\,\|u_{N_c} - u\|_{L^2}\left(\|u_{N_c} - u\|^2_{H^1} + N_c^{-1}\|u_{N_c} - u\|_{H^1}\right)$$

which completes the proof of Lemma 3.6. □

Using (2), we obtain, for all $\epsilon > 0$,

$$\|u - \Pi_{N_c} u\|_{H^1} \leq \frac{1}{N_c^{m+1/2-\epsilon}}\|u\|_{H^{m+1/2-\epsilon}}.$$

Therefore we deduce from lemma 3.6, that for all $\epsilon > 0$,

$$\|u_{N_c} - u\|_{H^s} \;\leq\; \frac{C}{{N_c}^{m+1/2-\epsilon-s}} \tag{85}$$

$$|\lambda_{N_c} - \lambda| \;\leq\; \frac{C}{{N_c}^{m+1/2-\epsilon}} \tag{86}$$

$$\frac{\gamma}{2}\|u_{N_c} - u\|^2_{H^1} \leq E(u_{N_c}) - E(u) \;\leq\; C\|u_{N_c} - u\|^2_{H^1},$$

for $s = 0$ and $s = 1$.
From (85) and the inverse inequality

$$\forall v_{N_c} \in V_{N_c}, \quad \|v_{N_c}\|_{H^r} \leq 2^{(r-s)/2}{N_c}^{r-s}\|v_{N_c}\|_{H^s}, \tag{87}$$

which holds true for all $s \leq r$ and all $N_c \geq 1$, we then obtain using classical arguments that, for all $\epsilon > 0$,

$$\|u_{N_c} - u\|_{H^s} \leq \frac{C}{{N_c}^{m+1/2-s-\epsilon}} \qquad \text{for all } 0 \leq s < m + 1/2 - \epsilon. \tag{88}$$

The estimate (86) is slightly deceptive since, in the case of a linear eigenvalue problem (i.e. for $-\Delta u + V^{\rm ion} u = \lambda u$) the convergence of the eigenvalues goes twice as fast as the convergence of the eigenvector in the H^1-norm. We are going to prove that this is also the case for the nonlinear eigenvalue problem. Let us first come back to (71), which we rewrite as,

$$\lambda_{N_c} - \lambda = \langle (A_{\rho^0} - \lambda)(u_{N_c} - u), (u_{N_c} - u)\rangle_{X',X} + \int_\Gamma w_{u,u_{N_c}}(u_{N_c} - u) \tag{89}$$

with

$$w_{u,u_{N_c}} = w^F_{u,u_{N_c}} + w^J_{u,u_{N_c}}$$

As $u/2 \leq u_{N_c} \leq 2u$ on Γ for N_c large enough, as u_{N_c} converges, hence is uniformly bounded, in $H^{m+1/2-\epsilon}_\#(\Gamma)$, for all $\epsilon > 0$, and since $H^{m+1/2-\epsilon}(\Gamma)$ is an algebra, $w^J_{u,u_{N_c}}$ and $w^F_{u,u_N}$are uniformly bounded in $H^{m+1/2-\epsilon}_\#(\Gamma)$ (at least for N_c large enough). We therefore infer from (89) that for N_c large enough

$$|\lambda_{N_c} - \lambda| \leq C\left(\|u_{N_c} - u\|^2_{H^1} + \|u_{N_c} - u\|_{H^{-m+3/2-\epsilon}}\right). \tag{90}$$

Let us now compute the H^{-r}-norm of the error for $r > 0$. Let $w \in H^r_\#(\Gamma)$. Proceeding as in Lemma 3.6 we obtain

$$\begin{aligned}
\int_\Gamma w(u_{N_c} - u) &= \langle (E''(u) - 2\lambda)(u_{N_c} - u), \Pi^1_{\widetilde{V}_{N_c}\cap u^\perp}\psi_w\rangle_{X',X} \\
&\quad + \langle (E''(u) - 2\lambda)(u_{N_c} - u), \psi_w - \Pi^1_{\widetilde{V}_N\cap u^\perp}\psi_w\rangle_{X',X} \\
&\quad + \|u_{N_c} - u\|^2_{L^2}\left[\int_\Gamma F''(u^2)u^3\psi_w + 2\sum_{k,\ell\in\mathcal{R}^*\setminus\{0\}} \frac{\widehat{(u^2)}_k\overline{\widehat{(u)}}_{k-\ell}\overline{\widehat{(\psi_w)}}_\ell}{|k|^2}\right] \\
&\quad - \frac{1}{2\mathcal{N}}\|u_{N_c} - u\|^2_{L^2}\int_\Gamma uw,
\end{aligned} \tag{91}$$

where $\Pi^1_{\widetilde{V}_{N_c}\cap u^\perp}$ denotes the orthogonal projector on $\widetilde{V}_{N_c} \cap u^\perp$ for the H^1 inner product. Then by noticing that ψ_w is in $H^{r+2}_\#(\Gamma)$ and satisfies

$$\|\psi_w\|_{H^{r+2}} \leq C\|w\|_{H^r}, \tag{92}$$

for some constant C independent of w, and combining it with (45), (82) - (84), (88), (89), (91) and (92), we obtain that there exists a constant $C \in \mathbb{R}_+$ such that for all $N_c \in \mathbb{N}$, and for all $\epsilon > 0$, and all $w \in H^r_\#(\Gamma)$,

$$\begin{aligned}
\int_\Gamma w(u_{N_c} - u) &\leq C'\left(\|u_{N_c} - u\|^2_{L^2} + {N_c}^{-(r+1)}\|u_{N_c} - u\|_{H^1}\right)\|w\|_{H^r} \\
&\leq \frac{C}{{N_c}^{m+1/2-\epsilon+r}}\|w\|_{H^r}.
\end{aligned}$$

Therefore, for all $\epsilon > 0$,

$$\|u_{N_c} - u\|_{H^{-r}} = \sup_{w \in H^r_{\#}(\Gamma)\setminus\{0\}} \frac{\int_\Gamma w(u_{N_c} - u)}{\|w\|_{H^r}} \leq \frac{C}{{N_c}^{m+1/2-\epsilon+r}}, \qquad (93)$$

for some constant $C \in \mathbb{R}_+$ independent of N_c. Using (88) and (90), we end up with

$$|\lambda_N - \lambda| \leq \frac{C}{{N_c}^{2m+1-\epsilon}}$$

which concludes the step 1 of the proof of the theorem

3.2 Step 2: proof of the uniqueness of u_{N_c}

Let focus on the proof of the uniqueness of u_{N_c}, for all $w_{N_c} \in V_{N_c}$, we have

$$\begin{aligned} E(w_{N_c}) - E(u_{N_c}) &= \langle A_{u^2_{N_c}} w_{N_c}, w_{N_c}\rangle_{X',X} - \langle A_{u^2_{N_c}} u_{N_c}, u_{N_c}\rangle_{X',X} \\ &\quad + C_{\mathrm{TF}} \int_\Gamma \left[F(|w_{N_c}|^2) - F(|u_{N_c}|^2) - F'(|u_{N_c}|^2)(w^2_{N_c} - u^2_{N_c})\right] \\ &\quad + 2\pi \sum_{k\in\mathcal{R}^*\setminus\{0\}} \frac{\left[\widehat{(w^2_{N_c})_k}\,\overline{\widehat{(w^2_{N_c})}}_k - \widehat{(u^2_{N_c})_k}\,\overline{\widehat{(u^2_{N_c})}}_k - 2\widehat{(u^2_{N_c})_k}\left\{\overline{\widehat{(w^2_{N_c})}} - \overline{\widehat{(u^2_{N_c})}}_k\right\}\right]}{|k|^2}. \end{aligned}$$

Since $w \mapsto F'(w)$ and $w \mapsto \displaystyle\sum_{k\in\mathcal{R}^*\setminus\{0\}} \frac{\widehat{w}_k\overline{\widehat{w}}_k}{|k|^2}$ are convex, then

$$\begin{aligned} &\int_\Gamma \left[F(|w_{N_c}|^2) - F(|u_{N_c}|^2) - F'(|u_{N_c}|^2)(w^2_{N_c} - u^2_{N_c})\right] \\ &\quad + 2\pi \sum_{k\in\mathcal{R}^*\setminus\{0\}} \frac{\left[\widehat{(w^2_{N_c})_k}\,\overline{\widehat{(w^2_{N_c})}}_k - \widehat{(u^2_{N_c})_k}\,\overline{\widehat{(u^2_{N_c})}}_k - 2\widehat{(u^2_{N_c})_k}\left\{\overline{\widehat{(w^2_{N_c})}} - \overline{\widehat{(u^2_{N_c})}}_k\right\}\right]}{|k|^2} \end{aligned}$$

is non negative. Therefore,

$$\begin{aligned} E(w_{N_c}) - E(u_{N_c}) &\geq \langle A_{u^2_{N_c}} w_{N_c}, w_{N_c}\rangle_{X',X} - \langle A_{u^2_{N_c}} u_{N_c}, u_{N_c}\rangle_{X',X} \\ &\geq \langle (A_{u^2_{N_c}} - \lambda_{N_c})(w_{N_c} - u_{N_c}), w_{N_c} - u_{N_c}\rangle_{X',X}. \quad (94) \end{aligned}$$

Let be μ_{N_c} the ground state eigenvalue of $A_{u^2_{N_c}}$ in V_{N_c} associated to w_{N_c}, we have that

$$\begin{aligned} \mu_{N_c} &= \inf_{w\in V_{N_c}} \frac{\langle A_{\rho_{N_c}} w, w\rangle_{X',X}}{(w,w)_{L^2}} \\ &= \inf_{w\in V_{N_c}} \frac{\langle A_{\rho^0} w, w\rangle_{X',X}}{(w,w)_{L^2}} + \frac{\langle (A_{u^2_{N_c}} - A_{\rho^0})w, w\rangle_{X',X}}{(w,w)_{L^2}} \qquad (95) \end{aligned}$$

Besides, we know from Lemma 3.6 that $(u_{N_c})_{N_c\in\mathbb{N}}$ converges to u in $H^{m+3/2-\epsilon}_{\#}(\Gamma)$ for $m>3$ and $\epsilon>0$, hence in $L^\infty_{\#}(\Gamma)$, when N_c goes to infinity. This implies that the sequence $(A_{\rho^0}-A_{u^2_{N_c}})_{N_c\in\mathbb{N}}$ converges to 0 in operator norm. Hence using this in (95) involves that μ_{N_c} converges to λ.
By contradiction we can prove that for N_c, large enough, the ground state eigenvalue μ_{N_c} of $A_{u^2_{N_c}}$ is simple and equal to λ_{N_c}. Indeed, supposing that there exists a sequence $(N_k)_k$ such that $\lambda_{N_k}\geq\mu_{N_k}$ and $w_{N_k}\neq\pm u_{N_k}$, then $(w_{N_k},u_{N_k})_{L^2}=0$. Therefore from (34), we get

$$\begin{aligned}\langle(A_{\rho^0}-\lambda)w_{N_k},w_{N_k}\rangle_{X',X} &\geq (\lambda_2-\lambda_1)\,(\|w_{N_k}\|^2_{L^2}-\frac{1}{\mathcal{N}}|\,(u,w_{N_k})|^2)\\ &\geq (\lambda_2-\lambda_1)\,(\mathcal{N}-\frac{1}{\mathcal{N}}|(u-u_{N_k},w_{N_k})|^2)\\ &\qquad\text{(by using the fact that } (w_{N_k},u_{N_k})_{L^2}=0)\\ &\geq (\lambda_2-\lambda_1)(\mathcal{N}-\|u-u_{N_k}\|^2_{L^2})\end{aligned}$$

Taking into account that $A_{u^2_{N_k}}$ converges to A_ρ as an operator and μ_{N_k} converges to λ, we obtain that

$$\begin{aligned}\langle(A_{u^2_{N_k}}-\mu_{N_k})w_{N_k},w_{N_k}\rangle_{X',X} &\underset{N_k\to+\infty}{\longrightarrow} \langle(A_{\rho^0}-\lambda)w_{N_k},w_{N_k}\rangle_{X',X}\\ &\geq(\lambda_2-\lambda_1)(\mathcal{N}-\|u-u_{N_k}\|^2_{L^2})\\ &\underset{N_k\to+\infty}{\longrightarrow}(\lambda_2-\lambda_1)\mathcal{N}>0\end{aligned}$$

which is impossible, since

$$\langle(A_{u^2_{N_k}}-\mu_{N_k})w_{N_k},w_{N_k}\rangle_{X',X}=0.$$

The fact that λ_{N_c} is the ground state eigenvalue of $A_{\rho_{N_c}}$ and is simple, implies that there exists a positive constant η_{N_c} such that

$$\langle(A_{u^2_{N_c}}-\lambda_{N_c})w,w\rangle_{X',X}\geq\eta_{N_c}(\|w\|^2_{L^2}+(\mathcal{N}-2)|(u_{N_c},w)_{L^2}|^2)>0\qquad\forall w\in V_{N_c}.$$

Using this in (94) we get that

$$E(w_{N_c})-E(u_{N_c})>0\qquad w_{N_c}\in V_{N_c}.$$

Therefore, for N_c large enough, u_{N_c} is the unique minimizer of (17).

3.3 Step 3: second part of the *a priori* errors estimates

Let us now turn to the pseudospectral approximation (17) of (16). First, we notice that

$$\begin{aligned}\frac{C_{\mathrm{W}}}{2}\|\nabla u_{N_c,N_g}\|^2_{L^2_{\#}}-\|V^{\mathrm{ion}}\|_{L^\infty}\mathcal{N} &\leq E^{\mathrm{TFW}}_{N_g}(u_{N_c,N_g})\\ &\leq E^{\mathrm{TFW}}_{N_g}(\mathcal{N}^{1/2}|\Gamma|^{-1/2})\\ &\leq C_{\mathrm{TF}}\mathcal{N}^{5/3}|\Gamma|^{-2/3}+\|V^{\mathrm{ion}}\|_{L^\infty}\mathcal{N},\end{aligned}$$

from which we infer that u_{N,N_g} is uniformly bounded in $H^1_\#(\Gamma)$. We then see that

$$\begin{aligned}\lambda_{N_c,N_g} &= \mathcal{N}^{-1}\bigg[\frac{C_{\rm W}}{2}\int_\Gamma |\nabla u_{N_c,N_g}|^2 + \int_\Gamma \mathcal{I}_{N_g}(V^{\rm ion}|u_{N_c,N_g}|^2 + f(|u_{N_c,N_g}|^2)|u_{N_c,N_g}|^2) \\ &\qquad + D_\Gamma(|u_{N_c,N_g}|^2,|u_{N_c,N_g}|^2)\bigg].\end{aligned}$$

Using (6), (11) and (41), we obtain that λ_{N,N_c} also is uniformly bounded. Now,

$$\begin{aligned}\Delta u_{N_c,N_g} &= 2C_{\rm W}^{-1}\Pi_{N_c}\left(\mathcal{I}_{N_g}\left(f(|u_{N_c,N_g}|^2)u_{N_c,N_g}\right)\right) + 2C_{\rm W}^{-1}\Pi_{N_c}\left(\mathcal{I}_{N_g}\left(V^{\rm ion}u_{N_c,N_g}\right)\right) \\ &\quad + 2C_{\rm W}^{-1}\Pi_{N_c}\left(V^{\rm Coulomb}_{|u_{N_c,N_g}|^2}u_{N_c,N_g}\right) - 2C_{\rm W}^{-1}\lambda_{N_c,N_g}u_{N_c,N_g}, \qquad (96)\end{aligned}$$

and we deduce from (4), (6) and (8) that

$$\begin{aligned}\left\|\Pi_{N_c}\left(\mathcal{I}_{N_g}\left(f(|u_{N_c,N_g}|^2)u_{N_c,N_g}\right)\right)\right\|_{L^2_\#} &\le \left(\int_\Gamma \left(\mathcal{I}_{N_g}(f(|u_{N_c,N_g}|^2))\right)^2|u_{N_c,N_g}|^2\right)^{1/2} \\ &= \left(\sum_{x\in\mathcal{G}_{N_g}\cap\Gamma}\left(\frac{L}{N_g}\right)^3 f(|u_{N_c,N_g}(x)|^2)^2|u_{N_c,N_g}(x)|^2\right)^{1/2} \\ &\le \frac{5}{3}C_{\rm TF}\|u_{N_c,N_g}\|^{1/3}_{L^\infty}\left(\sum_{x\in\mathcal{G}_{N_g}\cap\Gamma}\left(\frac{L}{N_g}\right)^3|u_{N_c,N_g}(x)|^4\right)^{1/2} \\ &= \frac{5}{3}C_{\rm TF}\|u_{N_c,N_g}\|^{1/3}_{L^\infty}\|u_{N_c,N_g}\|^2_{L^4_\#},\end{aligned}$$

and that

$$\begin{aligned}\|\Pi_{N_c}\left(\mathcal{I}_{N_g}\left(V^{\rm ion}u_{N_c,N_g}\right)\right)\|_{L^2_\#} &\le \left(\int_\Gamma \mathcal{I}_{N_g}(|V^{\rm ion}|^2|u_{N_c,N_g}|^2)\right)^{1/2} \\ &\le \|V^{\rm ion}\|_{L^\infty}\mathcal{N}^{1/2}.\end{aligned}$$

Besides, using (30),

$$\begin{aligned}\|\Pi_{N_c}\left(V^{\rm Coulomb}_{|u_{N_c,N_g}|^2}u_{N_c,N_g}\right)\|_{L^2_\#} &\le \|V^{\rm Coulomb}_{|u_{N_c,N_g}|^2}u_{N_c,N_g}\|_{L^2_\#} \\ &\le \mathcal{N}^{1/2}\|V^{\rm Coulomb}_{|u_{N_c,N_g}|^2}\|_{L^\infty} \\ &\le \mathcal{N}^{1/2}\|u_{N_c,N_g}\|^2_{L^4_\#}.\end{aligned}$$

As u_{N_c,N_g} is uniformly bounded in $H^1_\#(\Gamma)$, and therefore in $L^4_\#(\Gamma)$, we get

$$\begin{aligned}\|u_{N_c,N_g}\|_{H^2_\#} &= \left(\|u_{N_c,N_g}\|^2_{L^2_\#} + \|\Delta u_{N_c,N_g}\|^2_{L^2_\#}\right)^{1/2} \\ &\le C\left(1+\|u_{N_c,N_g}\|^{1/3}_{L^\infty}\right) \\ &\le C\left(1+\|u_{N_c,N_g}\|^{1/3}_{H^2_\#}\right).\end{aligned}$$

Therefore u_{N_c,N_g} is uniformly bounded in $H^2_\#(\Gamma)$, hence in $L^\infty(\mathbb{R}^3)$.
Returning to (96) and using (9) and a bootstrap argument, we conclude that u_{N_c,N_g} is in fact uniformly bounded in $H^{7/2+\epsilon}_\#(\Gamma)$.
Next, using (94),

$$\begin{aligned}
\frac{\gamma}{2}\|u_{N_c,N_g}-u_{N_c}\|^2_{H^1} &\leq E^{\mathrm{TFW}}(u_{N_c,N_g})-E^{\mathrm{TFW}}(u_{N_c})\\
&= E^{\mathrm{TFW}}_{N_g}(u_{N_c,N_g})-E^{\mathrm{TFW}}_{N_g}(u_{N_c})\\
&\quad+\int_\Gamma((1-\mathcal{I}_{N_g})(V))(|u_{N_c,N_g}|^2-|u_{N_c}|^2)\\
&\quad+\int_\Gamma(1-\mathcal{I}_{N_g})(F(|u_{N_c,N_g}|^2)-F(|u_{N_c}|^2))\\
&\leq \int_\Gamma((1-\mathcal{I}_{N_g})(V))(|u_{N_c,N_g}|^2-|u_{N_c}|^2)\\
&\quad+\int_\Gamma(1-\mathcal{I}_{N_g})(F(|u_{N_c,N_g}|^2)-F(|u_{N_c}|^2)).
\end{aligned}$$

Let $g(t,t')=\frac{F(t'^2)-F(t^2)}{t'-t}$. For N_c large enough, u_{N_c} is uniformly bounded away from zero; besides, both u_{N_c} and u_{N_c,N_g} are uniformly bounded in $H^{7/2+\epsilon}_\#(\Gamma)$. Therefore, $g(u_{N_c},u_{N_c,N_g})$ is uniformly bounded in $H^{7/2+\epsilon}_\#(\Gamma)$. This implies that the Fourier coefficients of $g(u_{N_c},u_{N_c,N_g})$ go to zero faster that $|k|^{-7/2}$, which implies, using (5) and (10), that

$$\begin{aligned}
&\left|\int_\Gamma(1-\mathcal{I}_{N_g})(F(|u_{N_c,N_g}|^2)-F(|u_{N_c}|^2))\right|\\
&\quad=\left|\int_\Gamma(1-\mathcal{I}_{N_g})\left(g(u_{N_c},u_{N_c,N_g})\right)\,(u_{N_c,N_g}-u_{N_c})\right|\\
&\quad\leq\left\|\Pi_{N_c}\left((1-\mathcal{I}_{N_g})\left(g(u_{N_c},u_{N_c,N_g})\right)\right)\right\|_{L^2_\#}\|u_{N_c,N_g}-u_{N_c}\|_{L^2_\#}\\
&\quad\leq CN_c^{3/2}N_g^{-7/2}\|u_{N_c,N_g}-u_{N_c}\|_{L^2_\#}. \qquad (97)
\end{aligned}$$

On the other hand,

$$\begin{aligned}
&\left|\int_\Gamma((1-\mathcal{I}_{N_g})(V))(|u_{N_c,N_g}|^2-|u_{N_c}|^2)\right|\\
&\quad\leq\|\Pi_{2N_c}((1-\mathcal{I}_{N_g})(V))\|_{L^2_\#}\|u_{N_c,N_g}+u_{N_c}\|_{L^\infty}\|u_{N_c,N_g}-u_{N_c}\|_{L^2_\#}\\
&\quad\leq CN_c^{3/2}N_g^{-m}\|u_{N_c,N_g}-u_{N_c}\|_{L^2_\#}.
\end{aligned}$$

Therefore,

$$\|u_{N_c,N_g}-u_{N_c}\|_{H^1_\#}\leq CN_c^{3/2}N_g^{-7/2}. \qquad (98)$$

We then deduce from (98) and the inverse inequality (87) that (u_{N_c,N_g}) converges to u in $H^2_\#(\Gamma)$, and therefore in $L^\infty(\mathbb{R}^3)$, when $N_g\geq 4N_c+1$. It follows that for N_c large enough, u_{N_c,N_g} is bounded away from zero, which, together

with (96), implies that (u_{N_c,N_g}) is bounded in $H^{m+1/2-\epsilon}_{\#}(\Gamma)$, when $N_g \geq 4N_c+1$. The estimates (97) and (98) can therefore be improved, yielding

$$\left|\int_\Gamma (1-\mathcal{I}_{N_g})(F(|u_{N_c,N_g}|^2)-F(|u_{N_c}|^2))\right| \leq C N_c^{3/2} N_g^{-(m+1/2-\epsilon)} \|u_{N_c,N_g}-u_{N_c}\|_{L^2_\#}.$$

and

$$\|u_{N_c,N_g}-u_{N_c}\|_{H^1_\#} \leq C N_c^{3/2} N_g^{-m}.$$

We deduce (23) from the inverse inequality (87). For N_c large enough, u_{N_c,N_g} is bounded away from zero, so that $f(|u_{N_c,N_g}|^2)$ is uniformly bounded in $H^{m+1/2-\epsilon}_{\#}(\Gamma)$. Therefore, the k^{th} Fourier coefficient of $(V^{\text{ion}}+f(|u_{N_c,N_g}|^2))$ is bounded by $C|k|^{-m}$ where the constant C does not depend on N_c and N_g. Using the equality

$$\begin{aligned}\lambda_{N_c,N_g}-\lambda_{N_c} &= \mathcal{N}^{-1}\Big[\langle (A_{|u_{N_c}|^2}-\lambda_{N_c})(u_{N_c,N_g}-u_{N_c}),(u_{N_c,N_g}-u_{N_c})\rangle_{H^{-1}_\#,H^1_\#} \\ &\quad -\int_\Gamma (1-\mathcal{I}_{N_g})(V^{\text{ion}}+f(|u_{N_c,N_g}|^2))|u_{N_c,N_g}|^2 \\ &\quad +D_\Gamma(|u_{N_c,N_g}|^2,|u_{N_c,N_g}|^2-|u_{N_c}|^2)+\int_\Gamma (f(|u_{N_c,N_g}|^2)-f(|u_{N_c}|^2))|u_{N_c,N_g}|^2\Big],\end{aligned}$$

(23) and (41), we obtain (24). A similar calculation leads to (25).

3.4 Step 4: proof of the uniqueness of u_{N_c,N_g}

Lastly, we have for all $v_{N_c} \in V_{N_c}$,

$$\begin{aligned}&E^{\text{TFW}}_{N_g}(v_{N_c})-E^{\text{TFW}}_{N_g}(u_{N_c,N_g}) \qquad (99)\\ &= \langle (\widetilde{H}_{u_{N_c,N_g}}-\lambda_{N_c,N_g})(v_{N_c}-u_{N_c,N_g}),(v_{N_c}-u_{N_c,N_g})\rangle_{H^{-1}_\#,H^1_\#} \\ &\quad +\frac{1}{2}D_\Gamma(|v_{N_c}|^2-|u_{N_c,N_g}|^2,|v_{N_c}|^2-|u_{N_c,N_g}|^2) \\ &\quad +\sum_{x\in\mathcal{G}_{N_g}\cap\Gamma}\left(\frac{L}{N_g}\right)^3\left(F(|v_{N_c}(x)|^2)-F(|u_{N_c}(x)|^2)-f(|u_{N_c}(x)|^2)(|v_{N_c}(x)|^2-|u_{N_c}(x)|^2)\right) \\ &\geq \langle (\widetilde{H}_{u_{N_c,N_g}}-\lambda_{N_c,N_g})(v_{N_c}-u_{N_c,N_g}),(v_{N_c}-u_{N_c,N_g})\rangle_{H^{-1}_\#,H^1_\#}. \qquad (100)\end{aligned}$$

As u_{N_c,N_g} converges to u in $H^2_\#(\Gamma)$, the operator $\widetilde{H}^{N_g}_{|u_{N_c,N_g}|^2}-H_{\rho^0}$ converges to zero in operator norm. Reasoning as in the proof of the uniqueness of u_{N_c}, we obtain that for N_c large enough and $N_g \geq 4N_c+1$, we have for all $v_{N_c} \in V_{N_c}$ such that $\|v_{N_c}\|_{L^2_\#}=\mathcal{N}^{1/2}$ and $(v_{N_c},u_{N_c})_{L^2_\#} \geq 0$,

$$\langle (\widetilde{H}_{u_{N_c,N_g}}-\lambda_{N_c,N_g})(v_{N_c}-u_{N_c,N_g}),(v_{N_c}-u_{N_c,N_g})\rangle_{H^{-1}_\#,H^1_\#} \geq \frac{\gamma}{2}\|v_{N_c}-u_{N_c,N_g}\|^2_{H^1_\#}.$$

Thus the uniqueness of u_{N_c,N_g} for N_c large enough, which conclude the proof of the theorem.

□

Chapitre 3

Schémas à deux grilles pour la résolution de problèmes aux valeurs propres non linéaires

Dans cette partie on s'intéresse à la résolution de problèmes aux valeurs propres non linéaires à l'aide d'un *schéma à deux grilles*. Cette méthode consiste à approcher la solution u d'un problème aux valeurs propres non linéaire, dans un premier temps, par la fonction u_H solution du problème aux valeurs propres non linéaire discret sur un espace de discrétisation grossier X_H. Puis à utiliser la solution grossière ainsi obtenue pour résoudre un problème aux valeurs propres linéarisé, ou même un problème linéarisé avec second membre, sur un espace de discretisation fin X_h, on appellera $u_h^{H,2g}$ cette solution.

On montrera que, si les pas d'espace grossier H et fin h sont choisit de façon adéquate, alors l'erreur $\|u-u_h^{H,2g}\|_{H^1}$ est du même ordre que $\|u-u_h\|_{H^1}$, où u_h n'est autre que la solution du problème aux valeurs propres non linéaire discret résolu directement sur X_h. Ceci repose sur le fait que la contribution de u_H à l'erreur est mesurée en norme $L^2(\Omega)$ et possède donc un ordre plus élevé que si elle était mesurée en norme $H^1(\Omega)$, comme on l'a vu dans le chapitre précédent.

Ce procédé permet de diminuer le temps de calcul de la solution de notre problème aux valeurs propres non linéaire, car le calcul de u_H et $u_h^{H,2g}$ est moins coûteux que celui de u_h. Dans la première étape : le calcul de u_H est clairement moins coûteux que celui de u_h puisque le coût est fonction croissante de la dimension de l'espace discret. Dans la seconde étape, la complexité du calcul est plus petite lorsque on choisit de résoudre un problème aux valeurs propres linéaire à la place du problème aux valeurs propres non linéaire sur X_h et encore mieux lorsque l'on résout qu'un problème linéaire avec second membre.

L'idée de *schéma à deux grilles* a d'abord été introduite par Xu [39, 55, 56] pour la résolution de problèmes elliptiques non symétriques et non linéaires. Elle a également été utilisée pour la résolution des équations Navier-Stokes [1, 20, 29, 54].
Pour résoudre un problème non linéaire, il est naturel d'utiliser un procédé itératif. Ici il consisterait, à chaque étape, à résoudre un problème aux valeurs propres linéarisé construit à partir de la solution de l'itération précédente. C'est dans cette optique que Xu et Zhou [12, 57, 58] se sont intéressés à l'application de *schémas à deux grilles* pour la résolution de problèmes aux valeurs propres linéaires. Le principe de leur méthode est de remplacer, à chaque étape du procédé itératif, la résolution d'un problème aux valeurs propres linéarisé sur une grille fine par celui sur une grille grossière suivit de la résolution d'un problème avec second membre sur une grille moyenne, puis une grille fine raffinée localement.
Leur méthode réduit la complexité et le temps de calcul tout en obtenant des résultats du même ordre de précision que ceux obtenus en résolvant le problème aux valeurs propres linéarisé directement sur la grille fine. Mais ce procédé reste itératif sur une grille fine, celui que l'on propose dans ce chapitre le sera seulement sur une grille grossière.

1 Introduction

On s'intéresse aux problèmes aux valeurs propres non linéaires issus de l'étude du problème de minimisation suivant :

$$I = \inf \left\{ E(v),\ v \in X,\ \int_\Omega v^2 = 1 \right\}. \tag{1}$$

La fonctionnelle d'énergie E est de la forme :

$$E(v) = \frac{1}{2} a(v,v) + \frac{1}{2} \int_\Omega F(v^2) \tag{2}$$

avec

$$a(u,v) = \int_\Omega (A\nabla u) \cdot \nabla v + \int_\Omega V uv. \tag{3}$$

Pour simplifier les notations on appelera X l'espace $H_0^1(\Omega)$ où Ω est un domaine à frontière régulière, ou bien un polygone convexe de $\mathbb{R}^d$, avec $d = 1$, 2 ou 3. On supposera également que :

- $A \in (L^\infty(\Omega))^{d\times d}$, $A(x)$ est symétrique pour tout $x \in \Omega$ et (4)
  $\exists\, \alpha > 0$ tel que $\xi^T A(x)\xi \geq \alpha |\xi|^2$, $\forall\, \xi \in \mathbb{R}^d$ et presque tout $x \in \Omega$ (5)

- $V \in L^p(\Omega)$ pour $p > 2$ (6)

- $F \in C^1([0,+\infty),\mathbb{R}) \cap C^2((0,\infty),\mathbb{R})$ et $F'' > 0$ sur $(0,+\infty)$ (7)
  $\exists\, 0 \leq q < 2,\ \exists\, C \in \mathbb{R}_+$ tels que $\forall\, t \geq 0,\ |F'(t)| \leq C(1+t^q)$ (8)

- $F''(t)t$ reste borné au voisinage de 0 (9)
  $\forall\, R > 0,\ \exists\, C \in \mathbb{R}_+$ tel que $\forall\ 0 < t_1 \leq R, \forall\, t_2 \in \mathbb{R}$:

$$\begin{aligned} \left|(F'(t_2^2) - F'(t_1^2))t_2^2\right| &\leq C|t_2 - t_1| && (10)\\ \left|F'(t_2^2)t_2 - F'(t_1^2)t_2 - 2F''(t_1^2)t_1^2(t_2 - t_1)\right| &\leq C|t_2 - t_1|^2. && (11) \end{aligned}$$

Pour simplifier les notations on prendra $f(t) = F'(t)$.

Dans le Chapitre 1, on a pu voir que ce problème de minimisation admettait exactement deux solutions u et $-u$, on notera u, la solution positive sur Ω. On rappelle également que u est aussi la solution de l'équation d'Euler-Lagrange :

$$\forall\, v \in X, \quad \langle E'(u) - \lambda u, v\rangle_{X',X} = 0, \tag{12}$$

pour $\lambda \in \mathbb{R}$ (multiplicateur de Lagrange associé à la contrainte $\|u\|_{L^2} = 1$). Alors l'équation (12) et la contrainte $\|u\|_{L^2} = 1$, définissent un problème aux valeurs propres non linéaire de la forme suivante :

$$\begin{cases} A_v v = \lambda v \\ \|v\|_{L^2} = 1 \end{cases} \tag{13}$$

où pour tout $v \in X$,

$$A_v = -\mathrm{div}(A\nabla\cdot) + V + f(v^2). \tag{14}$$

A_v, pour v fixé, est un opérateur auto-adjoint sur $L^2(\Omega)$, de domaine $H^2(\Omega)\cap X$. La plus petite valeur propre de l'opérateur linéaire A_u, λ, est simple. Elle est également la plus petite valeur propre de l'opérateur non linéaire A_v, et est associée à l'état fondamental u (voir le Chapitre 1).

Si le domaine Ω est à frontière régulière, ou bien un polygone convexe de $\mathbb{R}^d$, avec $d = 1$, 2, ou 3, si V vérifie l'hypothèse (6) et si F vérifie les hypothèses (8)-(11) alors $u \in H^2(\Omega)$. De plus si $V \in H^1(\Omega)$, on peut montrer que $u \in H^3(\Omega)$.

On introduit deux sous-espaces de $H_0^1(\Omega)$ de dimension finie, de type éléments finis $(K,\, P_K,\, \sum_K)$, notés X_H^k et X_h^k tels que :

- $X_\delta^k = \{v \in H_0^1(\Omega),\, \forall\, K \in \mathcal{T}_\delta,\, v_{|K} \in \mathbb{P}_k(K)\}$,
- $k = \{1; 2\}$,
- $\delta = H$ ou h et $H >> h$,
- $\mathcal{T}_h$ est une sous triangulation de $\mathcal{T}_H$.

Reformulons le problème (13) de façon suivante :
Trouver $u \in X$, $\|u\|_{L^2} = 1$ et $\lambda \in \mathbb{R}$ tels que

$$a(u, v) + \int_\Omega f(u^2)uv = \lambda \int_\Omega uv, \qquad \forall\, v \in X. \tag{15}$$

Ainsi son approximation de Galerkin sur X_δ^k s'écrit :
Trouver $(u_{\delta,k}, \lambda_{\delta,k}) \in X_\delta^k \times \mathbb{R}$ tel que :

$$a(u_{\delta,k}, v_{\delta,k}) + \int_\Omega f(u_{\delta,k}^2)u_{\delta,k}v_{\delta,k} = \lambda_{\delta,k} \int_\Omega u_{\delta,k}v_{\delta,k} \quad \forall\, v_{\delta,k} \in X_{\delta,k} \text{ et } \|u_{\delta,k}\|_{L^2(\Omega)} = 1. \tag{16}$$

Soit $u_{\delta,k}^0$ la solution du problème de minimisation discret sur X_δ^k :

$$I_\delta = \inf\{E(w_{\delta,k}) |\, w_{\delta,k} \in X_\delta^k,\ \|w_{\delta,k}\|_{L^2} = 1\}. \tag{17}$$

Il a été montré dans le Chapitre 1 que $u_{\delta,k}$, la solution de (16) est également un minimiseur de (17). Il est clair que $-u_{\delta,k}$ est aussi une solution de (16) associée à la même la valeur propre $\lambda_{\delta,k}$, on notera $u_{\delta,k}$ celle qui vérifie la propriété suivante : $\displaystyle\int_\Omega u\, u_{\delta,k} \geq 0$.
On rappelle un résultat important du Chapitre 1 : Sous certaines hypothèses, il existe $c \in \mathbb{R}_+$ et $\delta_0 \in \mathbb{R}_+$ tels que pour tout $0 < \delta \leq \delta_0$ et $k \in \{1, 2\}$ on ait :

$$\|u - u_{\delta,k}\|_X \leq c\delta^k \|u\|_{H^{k+1}(\Omega)} \tag{18}$$

$$\|u - u_{\delta,k}\|_{L^2(\Omega)} \leq c\delta^{k+1} \|u\|_{H^{k+1}(\Omega)} \tag{19}$$

$$|\lambda - \lambda_{\delta,k}| \leq c\delta^{2k} \|u\|_{H^{k+1}(\Omega)}. \tag{20}$$

Dans cette partie on ne traitera que le cas des élements finis de type $\mathbb{P}_1$, l'espace de discrétisation sera alors noté X_δ au lieu de X_δ^k.
On appelle $(u_H,\ \lambda_H)$ le couple solution du problème (16) sur X_H tel que $(u, u_H)_{L^2} \geq 0$. Ainsi en utilisant (18)-(20) avec $\delta = H$ et $k = 1$, on obtient les estimations d'erreur suivantes :

$$\begin{aligned} \|u - u_H\|_X &\leq cH\|u\|_{H^2(\Omega)} && (21)\\ \|u - u_H\|_{L^2(\Omega)} &\leq cH^2\|u\|_{H^2(\Omega)} && (22)\\ |\lambda - \lambda_H| &\leq cH^2\|u\|_{H^2(\Omega)}. && (23) \end{aligned}$$

Dans ce chapitre on s'intéressera à trois techniques utilisant des schémas à deux grilles pour approcher u schématisées de la façon suivante.

1. Sur une grille grossière

Résolution d'un problème aux valeurs propres non linéaire sur un espace grossier X_H
$$a(u_H, v) + \int_\Omega f(u_H^2)u_H v = \lambda_H \int_\Omega u_H v, \quad \forall v \in X_H$$

2. Sur une grille fine

Problème 1	Problème 2	Problème 3
Résolution d'un problème aux valeurs propres linéarisé sur un espace fin X_h	Résolution d'un problème avec second membre linéarisé sur un espace fin X_h	Résolution d'un problème avec second membre linéarisé sur un espace fin X_h
$a(u_h^H, v) + \int_\Omega f(u_H^2)u_h^H v = \lambda_h^H \int_\Omega u_h^H v \quad \forall v \in X_h$	$a(\tilde{u}_h^H, v) + \int_\Omega f(u_H^2)\tilde{u}_h^H v = \lambda_H \int_\Omega u_H v \quad \forall v \in X_h$	$a(\overline{u}_h^H, v) = -\int_\Omega f(u_H^2)u_H v + \lambda_H \int_\Omega u_H v \quad \forall v \in X_h.$

Pour simplifier les notations, on appelera b la forme bilinéaire définie par :

$$\begin{aligned} b(v, w) &= \langle (E''(u) - \lambda)v, w\rangle_{X',X} && (24)\\ &= a(v, w) + \int_\Omega f(u^2)vw + 2\int_\Omega f'(u^2)u^2 vw - \lambda \int_\Omega vw. && (25) \end{aligned}$$

D'après le Lemme 2.1 du Chapitre 1, elle est continue et coercive sur X . Soit

d la forme bilinéaire définie par :

$$\begin{aligned} d(v,w) &= \langle (A_u - \lambda)v, w\rangle_{X',X} && (26)\\ &= a(v,w) + \int_\Omega f(u^2)vw - \lambda \int_\Omega vw. && (27)\\ & && (28)\end{aligned}$$

Elle est continue et coercive sur $u^\perp = \{v \in X,\ \int_\Omega uv = 0\}$.
En effet, en utilisant le fait que λ, la plus petite valeur propre de A_u, est simple, on a la propriété suivante : il existe une constante $\eta > 0$, égale à $\lambda_2 - \lambda$, telle que

$$\forall v \in u^\perp \qquad \langle (A_u - \lambda)v, v\rangle_{X',X} \geq \eta \|v\|^2_{L^2(\Omega)} \geq 0.$$

Ensuite en raisonnant par l'absurde on montre qu'il existe une constante $\mu > 0$ telle que

$$\forall v \in u^\perp \quad \langle (A_u - \lambda)v, v\rangle_{X,X'} \geq \mu \|v\|^2_{H^1} \tag{29}$$

Montrons dans un premier temps que l'application $G : v \mapsto \langle (A_u - \lambda)v, v\rangle_{X,X'}$ est faiblement semi-continue inférieurement (s.c.i) dans X. Remarquons d'abord pour cela que pour toute fonction v et w dans X

$$|\langle (A_u - \lambda)v, w\rangle_{X',X}| \leq C\|v\|_{H^1}\|w\|_{H^1}.$$

En effet,

$$\begin{aligned} |\langle (A_u - \lambda)v, w\rangle_{X',X}| &\leq c\,[\|A\|_{L^\infty}\|\nabla v\|_{L^2}\|\nabla w\|_{L^2} + \|V\|_{L^p}\|v\|_{L^{2p'}}\|w\|_{L^{2p'}}\\ &\qquad + \int_\Omega |f(u^2)vw|] \qquad \text{où } p' = (1-p^{-1})^{-1}\\ &\leq c\,[\|A\|_{L^\infty}\|\nabla v\|_{L^2}\|\nabla w\|_{L^2} + \|V\|_{L^p}\|v\|_{L^{2p'}}\|w\|_{L^{2p'}}\\ &\qquad + \int_\Omega |(1+u^{2q})vw|] \qquad \text{avec } 0 \leq q < 2\\ &\qquad\qquad (\text{ en utilisant (8) })\\ &\leq c\,[\|A\|_{L^\infty}\|\nabla v\|_{L^2}\|\nabla w\|_{L^2} + \|V\|_{L^p}\|v\|_{L^{2p'}}\|w\|_{L^{2p'}}\\ &\qquad + \|v\|_{L^2}\|w\|_{L^2} + \|u\|_{L^{2q}}\|v\|_{L^{2q'}}\|v\|_{L^{2q'}}]\\ &\qquad\qquad \text{où } q' = (1-1/(2q))^{-1} \text{ et } 0 < q < 2\\ &\leq C\|v\|_{H^1}\|w\|_{H^1}. \end{aligned}$$

Ceci montre la continuité de l'application G sur X pour la topologie forte. On note ensuite qu' elle est quadratique et positive, ceci montre que G est convexe. Ainsi il vient la semi-continuité inférieure faible de l'application G dans X.
Supposons maintenant qu'il existe une suite $(v_n)_{n\geq 1} \in u^\perp$ telle que $\|v_n\|_{H^1} = 1$ et que $\lim_{n\to+\infty} \langle (A_u - \lambda)v_n, v_n\rangle_{X,X'} = 0$.

Alors en utilisant le fait que G est s.c.i. pour la topologie faible et ainsi que sa positivité on obtient

$$0 \leq \langle (A_u - \lambda)v, v\rangle_{X,X'} \leq \lim_{n\to 0} \langle (A_u - \lambda)v_n, v_n\rangle_{X,X'} = 0$$

et $v = 0$.

La suite $(v_n)_{n\geq 1}$ étant bornée dans $H^1(\Omega)$, on peut extraire une sous-suite $(v_{n_k})_{n_k\geq 1}$ qui converge faiblement dans $H^1(\Omega)$ vers v dans X.
De plus, en utilisant la compacité de $H^1(\Omega)$ dans $L^p(\Omega)$ (avec $1 \leq p < 6$ si $d = 3$ et $1 \leq p < +\infty$ si $d = 2$), on peut extraire une sous-suite $(v_{n_{k'}})_{n_{k'}\geq 1}$ qui converge fortement dans $L^p(\Omega)$ vers v'. Ainsi, par unicité de la limite, il résulte

$$\left\{ \begin{array}{lcll} v_{n_k} & \rightharpoonup & v & \text{dans } H^1(\Omega) \text{ faible} \\ v_{n_k} & \to & v & \text{dans } L^p(\Omega) \text{ fort.} \end{array} \right.$$

En particulier

$$v_{n_k}^2 \to v^2 \qquad \text{dans } L^2(\Omega) \text{ fort.} \tag{30}$$

On obtient ainsi,

$$\lim_{n_k\to+\infty} \left| \int_\Omega V[v^2 - v_{n_k}^2]\right| \leq \|V\|_{L^2} \lim_{n_k\to+\infty} \|v^2 - v_{n_k}^2\|_{L^2} = 0 \text{ (d'après (30) et 6),}$$

de la même façon, on a

$$\begin{array}{rcll} \lim_{n_k\to+\infty} \left| \int_\Omega f(u^2)[v^2 - v_{n_k}^2]\right| & \leq & C \lim_{n_k\to+\infty} \int_\Omega |(1+u^{2q})[v^2 - v_{n_k}^2]| & \text{avec } 0 < q < 2 \\ & \leq & C(1 + \|u\|_{L^2}^{2q}) \lim_{n_k\to+\infty} \|v^2 - v_{n_k}^2\|_{L^2} & \\ & \leq & C \lim_{n_k\to+\infty} \|v^2 - v_{n_k}^2\|_{L^2} = 0 \qquad \text{car } \|u\|_{L^2} = 1. & \end{array}$$

Ainsi en utilisant ces dernières lignes et $v = 0$, il vient

$$0 = \int_\Omega Vv^2 + \int_\Omega f(u^2)v^2 \leq \lim_{n_k\to+\infty} \inf \int_\Omega Vv_{n_k}^2 + \int_\Omega f(u^2)v_{n_k}^2.$$

A partir de cette dernière ligne, on obtient

$$\begin{array}{rcll} 0 = \lim_{n_k\to+\infty} \langle (A_u - \lambda)v_{n_k}, v_{n_k}\rangle_{X,X'} & \geq & \lim_{n_k\to+\infty} \inf \langle (A_u - \lambda)v_{n_k}, v_{n_k}\rangle_{X,X'} & \\ & \geq & 0 + \lim_{n_k\to+\infty} \inf \int_\Omega |\nabla v_{n_k}|^2 & \\ & \geq & \alpha \|v_{n_k}\|_{H^1}^2 \text{ (en utilisant 4)} & \end{array}$$

ce qui est impossible puisque $\|v_n\|_{H^1} = 1$. Ceci termine la démonstration par l'absurde et donne (29).

2 Résolution d'un problème linéarisé aux valeurs propres sur la grille fine

Dans cette partie, on s'intéresse au cas où dans la seconde étape de la méthode à deux grilles, on choisit de résoudre le problème aux valeurs propres linéaire suivant :

Problème 1 : Trouver $u_h^H \in X_h$, $\|u_h^H\|_{L^2(\Omega)} = 1$ et $\lambda_h^H \in \mathbb{R}$ tels que

$$a(u_h^H, v_h) + \int_\Omega f(u_H^2) u_h^H v_h = \lambda_h^H \int_\Omega u_h^H v_h \qquad \forall\, v_h \in X_h. \tag{31}$$

Théorème 2.1 *Si V vérifie l'hypothèse (7) et si F vérifie les hypothèses (8)-(11), alors il existe $c \in \mathbb{R}_+$ et $\delta_0 \in \mathbb{R}_+$ tels que pour tout $0 < h \leq \delta_0$ et $0 < H \leq \delta_0$ on ait :*

$$\|u - u_h^H\|_X \leq c[h + H^2]\|u\ \|_{H^2(\Omega)}. \tag{32}$$

Ceci est semblable à l'estimation (21) dès que l'on choisit $h \sim H^2$.

Démonstration Dans un premier temps on montrera que $\|u - u_h^H\|_X \underset{h\to 0}{\longrightarrow} 0$. Pour cela, considérons le problème de minimisation suivant :

$$I^H = \inf\{E^H(v),\, v \in X,\, \|v\|_{L^2(\Omega)} = 1\}, \tag{33}$$

$$\text{où} \quad E^H(v) = \frac{1}{2}a(v,v) + \frac{1}{2}\int_\Omega f(u_H^2)v^2. \tag{34}$$

Lemme 2.2 *Le problème de minimisation (33) admet une unique solution positive dans X.*

Démonstration (Détails de la preuve dans l'annexe 5)
Schéma de preuve :

- I. Existence

1. On se donne une suite minimisante (u_{*n}^H) du problème (33)

$$\begin{cases} u_{*n}^H \in X \\ \\ \|u_{*n}^H\|^2_{L^2(\Omega} = 1 \\ \\ E^H(u_{*n}^H)\downarrow_{n\to+\infty} I = \inf\{E^H(v),\, v \in X,\, \|v\|^2_{L^2(\Omega} = 1\} \end{cases}$$

2. On montre (u_{*n}^H) est bornée dans X.
3. On montre $u_{*n}^H \rightharpoonup u_*^H$ dans X.
4. On montre

$$\begin{cases} (u_{*n}^H) \text{ bornée dans } X \\ u_{*n}^H \rightharpoonup u_*^H \text{ bornée dans } X \end{cases} \quad \Longrightarrow \|u_*^H\|^2_{L^2(\Omega)} = 1 \text{ et } \exists (u_{*n_k}^H)/E^H(u_*^H) \leq \liminf E^H(u_{*n_k}^H)$$

5. u_*^H est solution de (33)

- II. Unicité

On introduit la fonction densité $\rho^H = (u_*^H)^2$ et l'ensemble auquel elle appartient

$$\mathcal{K} = \{\rho \in L^1(\Omega), \rho \geq 0, \sqrt{\rho} \in X, \int_\Omega \rho = 1\}.$$

ainsi que le problème de minimisation suivant :

$$\tilde{I}^H = \inf\{\mathcal{E}^H(\rho) = \frac{1}{2}\int_\Omega (\nabla\sqrt{\rho})^2 + \frac{1}{2}\int_\Omega V\rho + \frac{1}{2}\int_\Omega f(u_H^2)\rho, \quad \rho \in \mathcal{K}\}. \tag{35}$$

On remarquera que minimiser E^H sur X avec la contrainte $\|v\|_{L^2} = 1$ est équivalent à minimiser $\mathcal{E}^H$ sur $\mathcal{K}$. En effet si u_*^H solution de (33) il vient que ρ^H est solution de (35) et de même si ρ^H est solution de (35) alors u_*^H solution de (33). Il suffira pour terminer de montrer que la fonctionnelle $\mathcal{E}^H$ est strictement convexe sur $\mathcal{K}$.

□

Notons $u_*^H \in X$, la solution du problème (33), elle est également solution de l'équation d'Euler-Lagrange suivante

$$\langle A_{u_H} u_*^H, v\rangle_{X',X} = a(u_*^H, v) + \int_\Omega f(u_H^2) u_*^H v = \lambda_*^H \int_\Omega u_*^H v \qquad \forall\, v \in X \tag{36}$$

où $\lambda_*^H \in \mathbb{R}$ est le multiplicateur de Lagrange associé à la contrainte

$$\|u_*^H\|_{L^2(\Omega)} = 1.$$

On s'intéresse au problème de valeurs propres linéaire (36) issu du problème de minimisation (33) et son approximation de Galerkin sur l'espace X_δ (avec $\delta = H$ ou h) qui s'écrit :
Trouver $u_\delta^H \in X_\delta$, $\|u_\delta^H\|_{L^2(\Omega)} = 1$, et $\lambda_\delta^H \in \mathbb{R}$ tels que

$$\langle A_{u_H} u_\delta^H, v_\delta\rangle_{X',X} = a(u_\delta^H, v_\delta) + \int_\Omega f(u_H^2) u_\delta^H v_\delta = \lambda_\delta^H \int_\Omega u_\delta^H v_\delta \qquad \forall\, v_\delta \in X_\delta. \tag{37}$$

On remarque que lorsque $\delta = h$ ce problème est identique au problème 1, donné par (31), de même lorsque $\delta = H$ ce problème est semblable au problème (16)

Notons $u_h^H \in X_h$ la solution de (31) telle que $(u_h^H, u_*^H)_{L^2(\Omega)} \geq 0$, montrons

$$\|u - u_h^H\|_X \underset{\substack{h\to 0\\ H\to 0}}{\longrightarrow} 0. \tag{38}$$

On commencera par montrer qu'il existe $M^H > 0$ tel que pour $\delta = H$ ou h, on a

$$\langle (A_{u_H} - \lambda_*^H)(u_\delta^H - u_*^H), (u_\delta^H - u_*^H)\rangle_{X',X} \geq M^H \|u_\delta^H - u_*^H\|_X^2. \tag{39}$$

Pour cela on utilise d'abord le fait que λ_*^H la plus petite valeur propre de A_{u_H} est simple, et que donc il existe $\eta^H > 0$ égale à $\lambda_{*,2}^H - \lambda_*^H$ tel que

$$\forall\, v^\perp \in (u_*^H)^\perp \qquad \langle (A_{u_H} - \lambda_*^H) v^\perp, v^\perp \rangle_{X',X} \geq \eta^H \|v^\perp\|^2_{L^2(\Omega)}.$$

Soit $v \in X$, alors $[v - (u_*^H, v)_{L^2(\Omega)} u_*^H] \in (u_*^H)^\perp$, on obtient ainsi que

$$\begin{aligned} \langle (A_{u_H} - \lambda_*^H) v, v \rangle_{X',X} &\geq \eta^H [\|v - (u_*^H, v)_{L^2(\Omega)} u_*^H]\|^2_{L^2(\Omega)} \\ &\geq \eta^H [\|v\|^2_{L^2(\Omega)} - |(u_*^H, v)_{L^2(\Omega)}|^2] \\ &\qquad (\text{ en utilisant que } \|u_*^H\|^2_{L^2(\Omega)} = 1). \end{aligned}$$

En supposant que $(u_*^H, u_\delta^H) > 0$ et en utilisant que $|(u_*^H, u_\delta^H)_{L^2(\Omega)}| \leq 1$, il apparait

$$\begin{aligned} \|u_\delta^H\|^2_{L^2(\Omega)} - |(u_*^H, u_\delta^H)_{L^2(\Omega)}|^2 &= 1 - |(u_*^H, u_\delta^H)_{L^2(\Omega)}|^2 \\ &\geq 1 - (u_*^H, u_\delta^H)_{L^2(\Omega)} = \frac{1}{2}\|u_\delta^H - u_*^H\|^2_{L^2(\Omega)}, \end{aligned}$$

De ce fait, il résulte

$$\langle (A_{u_H} - \lambda_*^H)(u_\delta^H - u_*^H), u_\delta^H - u_*^H \rangle_{X',X} \geq \frac{\lambda_{*,2}^H - \lambda_*^H}{2} \|u_\delta^H - u_*^H\|^2_{L^2(\Omega)}. \quad (40)$$

Soit $v \in X$, posons $p' = (1 - p^{-1})^{-1}$, de façon à ce que $V \in L^p(\Omega)$, $p \geq 2$, alors $\displaystyle\int_\Omega V v^2 \leq c\, \|V\|_{L^p} \|v\|^2_{L^{2p'}} \leq \infty$. Il en découle

$$\begin{aligned} \langle (A_{u_H} - \lambda_*^H) v, v \rangle_{X',X} &\geq \int_\Omega (A\nabla v) \cdot \nabla v + \int_\Omega (V + f(u_H^2)) v^2 - \lambda_*^H \int_\Omega v^2 \\ &\geq \alpha \|\nabla v\|^2_{L^2(\Omega)} - \|V\|_{L^p(\Omega)} \|v\|^2_{L^{2p'(\Omega)}} + (f(0) - \lambda_*^H)\|v\|^2_{L^2(\Omega)} \\ &\quad \text{(en utilisant (5))} \\ &\geq \alpha \|\nabla v\|^2_{L^2(\Omega)} - \|V\|_{L^p(\Omega)} \|v\|^{2-3/p}_{L^2(\Omega)} \|v\|^{3/p}_{L^6(\Omega)} + (f(0) - \lambda_*^H)\|v\|^2_{L^2(\Omega)} \\ &\geq \alpha \|\nabla v\|^2_{L^2} - C_6^{3/p} \|V\|_{L^p(\Omega)} \|v\|^{2-3/p}_{L^2} \|v\|^{3/p}_{H^1(\Omega)} + (f(0) - \lambda_*^H)\|v\|^2_{L^2(\Omega)} \\ &\geq \frac{\alpha}{2}\|\nabla v\|^2_{L^2(\Omega)} \\ &\quad + \left(f(0) - \frac{3-2p}{2p} \left(\frac{3C_6^2 \|V\|^{2p/3}_{L^p(\Omega)}}{p\alpha} \right)^{3/(2p-3)} - \frac{\alpha}{2} - \lambda_*^H \right) \|v\|^2_{L^2(\Omega)}, \end{aligned}$$

où C_6 est la constante venant de l'inégalité de Sobolev suivante : $\forall v \in X$, on a $\|v\|_{L^6(\Omega)} \leq C_6 \|v\|_{H^1(\Omega)}$. Par conséquent, Il existe une constante positive β telle que

$$\forall v \in X \qquad \langle (A_{u_H} - \lambda_*^H) v, v \rangle_{X',X} \leq \frac{\alpha}{2}\|\nabla v\|^2_{L^2(\Omega)} - (\beta + \lambda_*^H)\|v\|^2_{L^2(\Omega)}. \quad (41)$$

Soit θ définit telle que $0 \leq \theta \leq \dfrac{\lambda_{*,2}^H - \lambda_*^H}{\lambda_{*,2}^H + \lambda_*^H + 2\beta} < 1$. Ainsi en combinant (40) et (41), il vient

$$\begin{aligned}
\langle (A_{u_H} - \lambda_*^H)(u_\delta^H - u_*^H), u_\delta^H - u_*^H \rangle_{X',X} & \\
&\geq \theta\, \frac{\alpha}{2} \|\nabla(u_\delta^H - u_*^H)\|^2_{L^2(\Omega)} - \theta\,(\beta + \lambda_*^H)\|u_\delta^H - u_*^H\|^2_{L^2(\Omega)} \\
&\quad + (1-\theta)\frac{\lambda_{*,2}^H - \lambda_*^H}{2}\|u_\delta^H - u_*^H\|^2_{L^2(\Omega)} \\
&\geq \theta\, \frac{\alpha}{2} \|\nabla(u_\delta^H - u_*^H)\|^2_{L^2(\Omega)} + \frac{\lambda_{*,2}^H - \lambda_*^H}{2}\|u_\delta^H - u_*^H\|^2_{L^2(\Omega)} \\
&\quad - \frac{\theta}{2}\,(2\beta + \lambda_*^H + \lambda_{*,2}^H)\|u_\delta^H - u_*^H\|^2_{L^2(\Omega)} \\
&\geq M_H \|u_\delta^H - u_*^H\|^2_{H^1(\Omega)}.
\end{aligned}$$

On obtient ainsi (39) avec $M_H = \max\left(\dfrac{\theta\alpha}{2}, \dfrac{\lambda_{*,2}^H - \lambda_*^H}{2} - \dfrac{\theta}{2}\,(2\beta + \lambda_*^H + \lambda_{*,2}^H)\right)$.

D'ailleurs on remarque que

$$E^H(u_\delta^H) - E^H(u_*^H) \quad \geq \quad M^H \|u_\delta^H - u_*^H\|_X^2.$$

En effet

$$\begin{aligned}
E^H(u_\delta^H) - E^H(u_*^H) &= \frac{1}{2}[a(u_\delta^H, u_\delta^H) - a(u_*^H, u_*^H) + \int_\Omega f(u_H^2)[(u_\delta^H)^2 - (u_*^H)^2]] \\
&= \frac{1}{2}[a(u_\delta^H - u_*^H, u_\delta^H - u_*^H) - 2a(u_*^H, u_\delta^H - u_*^H)] \\
&\quad + \frac{1}{2}\int_\Omega f(u_H^2)(u_\delta^H - u_*^H)^2 - \int_\Omega f(u_H^2) u_*^H (u_\delta^H - u_*^H) \\
&= \frac{1}{2}[a(u_\delta^H - u_*^H, u_\delta^H - u_*^H) + \int_\Omega f(u_H^2)(u_\delta^H - u_*^H)^2] \\
&\quad - \lambda_*^H \int_\Omega u_*^H (u_\delta^H - u_*^H) \\
&= \frac{1}{2}[a(u_\delta^H - u_*^H, u_\delta^H - u_*^H) + \int_\Omega f(u_H^2)(u_\delta^H - u_*^H)^2] \\
&\quad - \frac{1}{2}\lambda_*^H \int_\Omega (u_\delta^H - u_*^H)^2 \\
&\qquad \text{(en utilisant que } \|u_*^H\|_{L^2(\Omega)} = \|u_\delta^H\|_{L^2(\Omega)} = 1) \\
&= \frac{1}{2}\langle (A_{u_H} - \lambda_*^H)(u_\delta^H - u_*^H), (u_\delta^H - u_*^H)\rangle_{X,X'} \\
&\geq M^H \|u_\delta^H - u_*^H\|_X^2 \qquad \text{(en utilisant (39))}.
\end{aligned}$$

Maintenant introduisons $v_{*,\delta}^H \in X_\delta$ tel que

$$\|u_*^H - v_{*,\delta}^H\|_X = \inf_{v_\delta \in X_\delta} \|u_*^H - v_\delta\|_X \text{ et } \|u_*^H - v_{*,\delta}^H\|_X \underset{\delta \to 0}{\longrightarrow} 0.$$

On a alors, puisque u_δ^H minimise E^H sur X_δ

$$\|u_\delta^H - u_*^H\|_X^2 \leq \frac{1}{M^H}[E^H(u_\delta^H) - E^H(u_*^H)] \leq \frac{1}{M^H}[E^H(v_{*,\delta}^H) - E^H(u_*^H)] \underset{\delta\to 0}{\longrightarrow} 0 \tag{42}$$

Ce dernier point découlant de la continuité de E^H. On note également que

$$\|u - u_h^H\|_X \leq \|u - u_H\|_X + \|u_H - u_*^H\|_X + \|u_*^H - u_h^H\|_X$$

et que le couple (u_h^H, λ_h^H) (resp. (u_H, λ_H)) est solution de (37) dans X_h (resp. X_H).

Supposons que $(u_H, u_*^H)_{L^2(\Omega)} \geq 0$, alors en utilisant (42) avec $\delta = h$ pour estimer le terme $\|u_*^H - u_h^H\|_X$, et avec $\delta = H$ pour estimer le terme $\|u_H - u_*^H\|_X$, ainsi que (21) avec $\delta = H$ pour estimer le terme$|u - u_H\|_X$, on retrouve (38).

Muni de cette convergence vers 0 on peut maintenant montrer que

$$\|u - u_h^H\|_X \leq c[h + H^2]\|u\|_{H^2(\Omega)}.$$

Soit $v_h \in X_h$, commençons par regarder le terme $b(v_h, u_h^H - u)$

$$\begin{aligned}
b(v_h, u_h^H - u) &= \langle (E''(u) - \lambda)v_h, u_h^H - u\rangle_{X',X} \\
&= \langle (E''(u) - \lambda)(u_h^H - u), v_h\rangle_{X',X} \\
&= a(u_h^H - u, v_h) + \int_\Omega f(u^2)(u_h^H - u)v_h + 2\int_\Omega f'(u^2)u^2(u_h^H - u)v_h \\
&\quad -\lambda \int_\Omega (u_h^H - u)v_h \\
&= a(u_h^H, v_h) + \int_\Omega f(u^2)u_h^H v_h - a(u, v_h) - \int_\Omega f(u^2)uv_h \\
&\quad +2\int_\Omega f'(u^2)u^2(u_h^H - u)v_h - \lambda\int_\Omega (u_h^H - u)v_h \\
&= \int_\Omega [f(u^2)u_h^H - f(u_H^2)u_h^H + 2f'(u^2)u^2(u_h^H - u)]v_h \\
&\quad -(\lambda_h^H - \lambda)\int_\Omega u_h^H v_h \\
&\qquad \text{(en utilisant (15) et (31))} \\
&= \int_\Omega [w^H u_h^H(u - u_H) + 2f'(u^2)u^2(u_h^H - u)]v_h \\
&\quad -(\lambda_h^H - \lambda)\int_\Omega u_h^H v_h
\end{aligned} \tag{43}$$

avec

$$w^H = (u + u_H)\frac{f(u^2) - f(u_H^2)}{u^2 - u_H^2}. \tag{44}$$

Commençons par montrer que u_H est également borné.
Soit $\mathcal{I}_H$ l'opérateur d'interpolation dans X_H.

Lemme 2.3 *(voir [4]) Pour tout entier l, $0 \leq l \leq k+1$, et pour tout r et q, $1 \leq r \leq q \leq +\infty$, tels que $W^{l,r}(K_H)$ soit inclus dans $\mathcal{C}^0(K_H)$, il existe une constante c positive ne dépendant que de l, r et q telle que, pour toute fonction v de $W^{l,r}(\Omega)$, on ait :*

$$\|v - \mathcal{I}_H v\|_{L^q(\Omega)} \leq cH^{l-\frac{d}{r}+\frac{d}{q}} |v\ |_{W^{l,r}(\Omega)} \tag{45}$$

Lemme 2.4 *(voir [4]) Pour tout entier l, $0 \leq l \leq k+1$, et pour tout r et q, $1 \leq r \leq q \leq +\infty$, tels que $W^{l,r}(K_H)$ soit inclus dans $\mathcal{C}^0(K_H)$ et dans $W^{1,q}(K_H)$, il existe une constante c positive ne dépendant que de l, r et q telle que, pour toute fonction v de $W^{l,r}(\Omega)$, on ait :*

$$\|v - \mathcal{I}_H v\|_{W^{1,q}(\Omega)} \leq cH^{l-1-\frac{d}{r}+\frac{d}{q}} |v\ |_{W^{l,r}(\Omega)} \tag{46}$$

En utilisant (45) avec $q = +\infty$, et $r = l = 2$ on obtient que :

$$\|u - \mathcal{I}_H u\|_{L^\infty(\Omega)} \leq cH^{2-\frac{d}{2}} |u\ |_{H^2(\Omega)}$$

alors pour $d = 2$ ou 3, on a

$$\|u - \mathcal{I}_H u\|_{L^\infty(\Omega)} \underset{H\to 0}{\longrightarrow} 0. \tag{47}$$

De la même façon, en utilisant (46), avec $l = q = r = 2$ on obtient que :

$$\|u - \mathcal{I}_H u\|_{H^1(\Omega)} \leq cH|u\ |_{H^2(\Omega)}. \tag{48}$$

Lemme 2.5 *(voir [15]) Il existe une constante c positive indépendante de H telle que, pour tout v_H dans X_H on ait :*

$$\|v_H\|_{L^\infty(\Omega)} \leq c\rho(H)\|v_H\|_{H^1(\Omega)} = \begin{cases} c(1+|\log H|)\|v_H\|_{H^1(\Omega)} & \text{si } d = 2 \\ cH^{-\frac{1}{2}}\|v_H\|_{H^1(\Omega)} & \text{si } d = 3. \end{cases} \tag{49}$$

Ainsi en utilisant le Lemme 2.5 on trouve :

$$\begin{aligned} \|u_H - \mathcal{I}_H u\|_{L^\infty(\Omega)} &\leq c\rho(H)\|u_H - \mathcal{I}_H u\|_{H^1(\Omega)} \\ &\leq c\rho(H)[\|u_H - u\|_{H^1(\Omega)} + \|u - \mathcal{I}_H u\|_{H^1(\Omega)}] \\ &\leq cH\rho(H)\|u\|_{H^2(\Omega)} \underset{H\to 0}{\longrightarrow} 0 \text{ (en utilisant (48) et (21)).} \end{aligned}$$

et donc, il en resulte

$$\|u_H\|_{L^\infty(\Omega)} < +\infty. \tag{50}$$

En utilisant maintenant le théorème des accroissements finis, il vient que $\dfrac{f(u^2) - f(u_H^2)}{u^2 - u_H^2}$ est borné. On a également que $u \in H^2(\Omega)$, et donc borné dans $L^\infty(\Omega)$. Ainsi en utilisant ceci et (50) dans (44), Il vient que

$$w^H = (u + u_H)\frac{f(u^2) - f(u_H^2)}{u^2 - u_H^2} \in L^\infty(\Omega). \tag{51}$$

Ainsi en revenant dans (43) on a que pour tout $v_h \in X_h$,

$$|b(v_h, u_h^H - u)| \leq c\|v_h\|_{L^2(\Omega)}[\|u - u_H\|_{L^2(\Omega)} + \|u - u_h^H\|_{L^2(\Omega)} + \ |\lambda_h^H - \lambda\|] \quad (52)$$

Il nous faut maintenant évaluer l'erreur commise sur les valeurs propres, et celle commise sur les fonctions propres évaluée en norme $L^2(\Omega)$.

$$\begin{aligned}
\lambda_h^H - \lambda &= \langle E'(u_h^H), u_h^H\rangle_{X',X} - \langle E'(u), u\rangle_{X',X} \\
&= a(u_h^H, u_h^H) - a(u,u) + \int_\Omega f(u_H^2)(u_h^H)^2 - \int_\Omega f(u^2)u^2 \\
&= a(u_h^H - u, u_h^H - u) + 2a(u, u_h^H - u) + \int_\Omega f(u_H^2)(u_h^H)^2 - \int_\Omega f(u^2)u^2 \\
&= a(u_h^H - u, u_h^H - u) + 2\lambda \int_\Omega u(u_h^H - u) - 2\int_\Omega f(u^2)u(u_h^H - u) \\
&\quad + \int_\Omega f(u_H^2)(u_h^H)^2 - \int_\Omega f(u^2)u^2 \\
&= a(u_h^H - u, u_h^H - u) - \lambda\|u_h^H - u\|_{L^2}^2 - 2\int_\Omega f(u^2)u(u_h^H - u) \\
&\quad + \int_\Omega f(u_H^2)(u_h^H)^2 - \int_\Omega f(u^2)u^2 \quad \text{(en utilisant que } \int_\Omega u^2 = \int_\Omega (u_h^H)^2 = 1) \\
&= a(u_h^H - u, u_h^H - u) - \lambda\|u_h^H - u\|_{L^2}^2 + \int_\Omega f(u^2)(u - u_h^H)^2 \\
&\quad + \int_\Omega (u_h^H)^2[f(u_H^2) - f(u^2)] \\
&= \langle (A_u - \lambda)(u_h^H - u), (u_h^H - u)\rangle_{X',X} + \int_\Omega w^H(u_H - u)(u_h^H)^2.
\end{aligned}$$

D'où

$$|\lambda - \lambda_h^H| \leq c[\|u - u_h^H\|_X^2 + \|u - u_H\|_{L^2(\Omega)}]. \quad (53)$$

En utilisant (22) avec $\delta = H$ et $k = 1$, on peut énoncer le lemme suivant :

Lemme 2.6 *Si V vérifie l'hypothèse (7) et si F vérifie les hypothèses (8)-(11), alors il existe $c \in \mathbb{R}_+$ et $\delta_0 \in \mathbb{R}_+$ tels que pour tout $0 < h \leq \delta_0$ et $0 < H \leq \delta_0$ on ait :*

$$|\lambda - \lambda_h^H| \leq c\|u - u_h^H\|_X^2 + H^2\|u\|_{H^2(\Omega)}]. \quad (54)$$

Regardons maintenant l'estimation de l'erreur en norme L^2, pour cela comme dans le Chapitre 1 on considère le problème adjoint suivant :
Trouver $\psi \in u^\perp$ tel que pour tout $v \in u^\perp$alors,

$$d(\psi, v) = \langle (A_u - \lambda)\psi, v\rangle_{X',X} = \int_\Omega (u - u_{Hh})v. \quad (55)$$

En utilsant la coercivité de la forme bilinéaire d sur $u^\perp$ et le Théorème de Lax-Milgram, on en déduit qu'il existe une unique solution au problème (55) et qu'elle vérifie les hypothèses de régularité suivantes :

$$\psi \in H_0^1(\Omega) \cap H^2(\Omega)$$

et

$$\|\psi\|_{H^2(\Omega)} \leq c\| \ u - u_h^H\|_{L^2(\Omega)}. \tag{56}$$

Ainsi il existe $\psi_h \in X_h$ telle que

$$\|\psi - \psi_h\|_{L^2(\Omega)} \leq ch^2\|\psi\|_{H^2(\Omega)} \leq ch^2\| \ u - u_h^H\|_{L^2(\Omega)} \tag{57}$$

et

$$\|\psi - \psi_h\|_X \leq ch\|\psi\|_{H^2(\Omega)} \leq ch\| \ u - u_h^H\|_{L^2(\Omega)} \tag{58}$$

Soit $u_h^{H*} \in X$ défini par

$$u_h^{H*} = u_h^H + (1 - \int_\Omega u u_{Hh})u, \tag{59}$$

de sorte que $u_h^{H*} - u \in u^\perp$, on remarque que

$$u_h^{H*} - u_h^H = \frac{1}{2} \ u\|u - u_h^H\|^2_{L^2(\Omega)}. \tag{60}$$

Alors

$$\begin{aligned}
\|u - u_h^H\|^2_{L^2} &= \int_\Omega (u - u_h^H)(u - u_h^{H*}) + \int_\Omega (u - u_h^H)(u_h^{H*} - u_h^H) \\
&= \int_\Omega (u - u_h^H)(u - u_h^{H*}) - \frac{1}{2}\|u - u_h^H\|^2_{L^2(\Omega)} \int_\Omega u(u - u_h^H) \\
&= \int_\Omega (u - u_h^H)(u - u_h^{H*}) - \frac{1}{2}\|u - u_h^H\|^2_{L^2(\Omega)} \int_\Omega u[(u - u_h^{H*}) + (u_h^{H*} - u_h^H)] \\
&= \int_\Omega (u - u_h^H)(u - u_h^{H*}) + \frac{1}{4}\|u - u_h^H\|^4_{L^2(\Omega)} \\
&= \langle (A_u - \lambda)\psi, u - u_h^{H*}\rangle_{X',X} + \frac{1}{4}\|u - u_h^H\|^4_{L^2(\Omega)} \\
&= \langle (A_u - \lambda)\psi, u - u_h^H\rangle_{X',X} + \frac{1}{4}\|u - u_h^H\|^4_{L^2(\Omega)} \\
&= \langle (A_u - \lambda)\psi_h, u - u_h^H\rangle_{X',X} + \langle (A_u - \lambda)(\psi - \psi_h), u - u_h^H\rangle_{X',X} \\
&\quad + \frac{1}{4}\|u - u_h^H\|^4_{L^2(\Omega)}
\end{aligned}$$

$$\begin{aligned}
\|u-u_h^H\|_{L^2}^2 &= \langle (A_u-\lambda)(u-u_h^H),\psi_h\rangle_{X',X}+\langle (A_u-\lambda)(\psi-\psi_h),u-u_h^H\rangle_{X',X}\\
&\quad +\frac{1}{4}\|u-u_h^H\|_{L^2(\Omega)}^4\\
&= -\langle (A_u-\lambda)u_h^H,\psi_h\rangle_{X',X}+\langle (A_u-\lambda)(\psi-\psi_h),u-u_h^H\rangle_{X',X}\\
&\quad +\frac{1}{4}\|u-u_h^H\|_{L^2(\Omega)}^4\\
&= \int_\Omega \left[f(u_H^2)-f(u^2)\right]u_h^H\psi_h-\langle (A_{u_H}-\lambda)u_h^H,\psi_h\rangle_{X',X}\\
&\quad +\langle (A_u-\lambda)(\psi-\psi_h),u-u_h^H\rangle_{X',X}+\frac{1}{4}\|u-u_h^H\|_{L^2(\Omega)}^4\\
&= \int_\Omega \left[f(u_H^2)-f(u^2)\right]u_h^H\psi_h+(\lambda-\lambda_{Hh})\int_\Omega u_h^H\psi_h\\
&\quad +\langle (A_u-\lambda)(\psi-\psi_h),u-u_h^H\rangle_{X',X}+\frac{1}{4}\|u-u_h^H\|_{L^2(\Omega)}^4\\
&= \int_\Omega [u_H-u]w^H u_h^H\psi_h+(\lambda-\lambda_{Hh})\int_\Omega u_h^H(\psi_h-\psi)\\
&\quad +(\lambda-\lambda_{Hh})\int_\Omega (u_h^H-u)\psi+\langle (A_u-\lambda)(\psi-\psi_h),u-u_h^H\rangle_{X',X}\\
&\quad +\frac{1}{4}\|u-u_h^H\|_{L^2(\Omega)}^4
\end{aligned}$$

en utilisant $\displaystyle\int_\Omega \psi u=0$.

En remplacant $|\lambda-\lambda_h^H|$ par son estimation obtenue dans le lemme 2.6 et en utilisant (57), (58) et (51) , on trouve que

$$\|u-u_h^H\|_{L^2(\Omega)}^2\le c[H^2\|u\|_{H^2(\Omega)}+h\|u-u_h^H\|_X]\|u-u_h^H\|_{L^2(\Omega}$$

On en déduit que

$$\|u-u_h^H\|_{L^2(\Omega)}\le c[H^2\|u\|_{H^2(\Omega)}+h\|u-u_h^H\|_X]. \tag{61}$$

En utilisant (52), (53) et (61), on trouve que pour tout $v_h\in X_h$,

$$\begin{aligned}
|b(v_h,u-u_h^H)| &\le c\|v_h\|_{L^2(\Omega)}[H^2\|u\|_{H^2(\Omega)}+\|u-u_h^H\|_X^2+h\|u-u_h^H\|_X]\\
&\le c\|v_h\|_X[H^2\|u\|_{H^2(\Omega)}+\|u-u_h^H\|_X^2+h\|u-u_h^H\|_X].
\end{aligned} \tag{62}$$

Passons maintenant à l'estimation de $\|u-u_h^H\|_X$. On peut montrer le Théorème 2.1 de deux façons, selon que l'on utilise la coercivité de b sur X ou sur le sous espace $u^\perp=\{v\in X,\displaystyle\int_\Omega uv=0\}$.

- La méthode la plus naturelle est d'utiliser la coercivité de b sur X. Pour cela on choisira $v_h=x_h-u_h^H$ où x_h est tel que

$$\|u-x_h\|_X=\inf_{v_h\in X_h}\|u-v_h\|_X\le ch\|u\|_{H^2(\Omega)}.$$

On note que

$$\|u - u_h^H\|_X \leq \|u - x_h\|_X + \|x_h - u_h^H\|_X. \tag{63}$$

D'après la coercivité de b sur X, il existe une constante $M > 0$ telle que

$$M\|x_h - u_h^H\|_X^2 \leq b(x_h - u_h^H, x_h - u) + b(x_h - u_h^H, u - u_h^H).$$

De plus en utilisant (62), on obtient

$$\begin{aligned} M\|x_h - u_h^H\|_X^2 \leq & b(x_h - u_h^H, x_h - u)) + c[H^2\|u\|_{H^2(\Omega)} + h\|u - u_h^H\|_X]\|x_h - u_h^H\|_X \\ \leq & c\|x_h - u_h^H\|_X[\|u - x_h\|_X + H^2\|u\|_{H^2(\Omega)} + h\|u - u_h^H\|_X]. \end{aligned}$$

Finalement, à partir de (63) on trouve que

$$\|u - u_h^H\|_X \leq c\|u - x_h\|_X + cH^2\|u\|_{H^2(\Omega)} \leq c[h + H^2]\|u\|_{H^2(\Omega)}. \tag{64}$$

- Une autre méthode pour estimer $\|u - u_h^H\|_X$ est d'utiliser la coercivité de b sur $u^\perp$. Le même raisonnement pourra alors être employé dans le cas des équations de Kohn-Sham.

On note que

$$\|u - u_h^H\|_X \leq \|u - u_h^{H*}\|_X + \|u_h^{H*} - u_h^H\|_X \tag{65}$$

où u_{Hh}^* est définit par (59).

Pour simplifier les notations, on appelera $v = u - u_h^{H*}$ et l'on choisira v_h telle que

$$v_h \quad = \quad \Pi_h v = \Pi_h(u - u_h^{H*})$$

où Π_h est l'opérateur de projection dans X_h muni du produit scalaire H^1, ainsi

$$\|v - v_h\|_X \leq \|v\|_X + \|v_h\|_X \leq c\|v\|_X \tag{66}$$

et

$$\begin{aligned} v - v_h &= v - \Pi_h v \\ &= u - u_h^H - u + u\int_\Omega u u_h^H - \Pi_h(u - u_h^H - u + u\int_\Omega u u_h^H) \\ &= -u_h^H + u\int_\Omega u u_h^H - \Pi_h(-u_h^H + u\int_\Omega u u_h^H) \\ &= -u_h^H + u\int_\Omega u u_h^H + u_h^H - \Pi_h u\int_\Omega u u_h^H \\ &= (u - \Pi_h u)\int_\Omega u u_h^H \end{aligned}$$

on obtient alors

$$\|v - v_h\|_X \leq c\|u - \Pi_h u\|_X \leq ch^{r-1}\|u\|_{H^r(\Omega)} \quad 1 \leq r \leq 2. \tag{67}$$

La forme bilinéaire b étant coercive sur $u^{\perp}$, il existe donc un $M > 0$ tel que

$$\begin{aligned} M\|v\|_X^2 &\leq b(v,v) \\ &\leq b(v-v_h,v) + b(v_h,v) \\ &\leq b(v-v_h,v) + b(v_h,u_h^{H*} - u_h^H) + b(v_h,u_h^H - u) \\ &\leq b(v-v_h,v) + \frac{1}{2}\,\|u-u_h^H\|_{L^2(\Omega)}^2 b(v_h,u) + b(v_h,u_h^H - u) \\ &\qquad \text{(en utilisant (60)).} \end{aligned}$$

En utilisant (66) on obtient

$$|b(v_h,u-u_h^H)| \leq c\|v\|_X[H^2\|u\|_{H^2(\Omega)} + \|u-u_h^H\|_X^2 + h\|u-u_h^H\|_X] \tag{68}$$

Regardons maintenant le terme $b(v-v_h,v)$. On a

$$|b(v-v_h,v)| \quad \leq \quad c\|v-v_h\|_X\ \|v\|_X$$

et en utilsant (67) avec $r=2$, on obtient donc,

$$|b(v-v_h,v)| \leq ch\|u\|_{H^2(\Omega)}\|v\|_X. \tag{69}$$

Il reste alors à évaluer le terme $b(v_h,u)$. On a

$$|b(v_h,u)| \leq c\|u\|_X\|v\|_X. \tag{70}$$

Finalement en regroupant (68) - (70), on trouve que

$$\begin{aligned} M\|v\|_X^2 &\leq c\|v\|_X[h\|u\|_{H^2(\Omega)} + \|u-u_h^H\|_{L^2(\Omega)}^2 + H^2\|u\|_{H^2(\Omega)} \\ &\qquad +\|u-u_h^H\|_X^2 + h\|u-u_h^H\|_X] \\ &\leq c\|v\|_X[(h+H^2)\|u\|_{H^2(\Omega)} + \|u-u_h^H\|_X^2 + h\|u-u_h^H\|_X] \end{aligned}$$

et donc que

$$\|v\|_X \leq \frac{c}{M}[(h+H^2)\|u\|_{H^2(\Omega)} + \|u-u_h^H\|_X^2 + h\|u-u_h^H\|_X]. \tag{71}$$

A partir de cette dernière inégalité dans (65), on obtient que

$$\|u-u_h^H\|_X \quad \leq \quad C[(h+H^2)\|u\|_{H^2(\Omega)} + \|u-u_h^H\|_X^2 + h\|u-u_h^H\|_X]$$

Finalement, en utilisant (38) on trouve que

$$\|u-u_h^H\|_X \quad \leq \quad C[h+H^2]\|u\|_{H^2(\Omega)},$$

dès que h est assez petit, ce qui termine la démonstration du théorème 2.1. □

3 Résolution d'un problème linéarisé avec second membre sur la grille fine

Dans cette partie, on traitera le cas où l'on choisit de résoudre un problème linéaire avec second membre dans la seconde étape de la méthode à deux grilles. Deux possibilités s'offrent à nous. La première consiste à résoudre le problème suivant, que l'on notera

Problème 2 :

Trouver $\tilde{u}_h^H \in X_h$ telle que

$$\forall\, v_h \in X_h, \qquad a(\tilde{u}_h^H, v_h) + \int_\Omega f(u_H^2)\tilde{u}_h^H v_h = \lambda_H \int_\Omega u_H v_h. \tag{72}$$

L' alternative est de résoudre le problème suivant, que l'on notera

Problème 3 : Trouver $\overline{u}_h^H \in X_h$ telle que

$$\forall\, v_h \in X_h, \qquad a(\overline{u}_h^H, v_h) = \lambda_H \int_\Omega u_H v_h - \int_\Omega f(u_H^2) u_H v_h. \tag{73}$$

Pour simplifier les notations, on appelera g la forme bilinéaire définie par :

$$g(w, v) = a(w, v) + \int_\Omega f(u_H^2) wv \qquad \forall\, w, v \in X. \tag{74}$$

De plus on supposera que V et F sont tels que la forme bilinéaire g vérifie la propritété suivante :

$$g(v, v) \geq \xi \|v\|_X^2 \quad \forall\, v \in X. \tag{75}$$

Commençons par montrer que les formes bilinéaires g et a sont continues dans X. D'après l'hypothèse (6) on a,

$$|\int_\Omega Vvw| \leq c\, \|v\|_X \|w\|_X \tag{76}$$

en effet

$$\begin{aligned} |\int_\Omega Vvw| &\leq \int_\Omega |Vvw| \\ &\leq \|V\|_{L^p(\Omega)} \|v\|_{L^{2p'}(\Omega)} \|w\|_{L^{2p'}(\Omega)} \qquad (\text{ où } p' = \frac{p}{p-1}) \end{aligned}$$

et $H^1(\Omega) \hookrightarrow L^{2p'}(\Omega)$ dès lors que $p \geq \frac{3}{2}$. Ainsi en combinant ceci avec (5), il vient

$$\forall\, v,\, w \in X \quad |a(v, w) \leq c \|v\|_X \|w\|_X. \tag{77}$$

Par ailleurs, en utilisant l'hypothèse (8), on obtient

$$\begin{aligned} \forall\, v,\, w \in X \qquad |\int_\Omega f(u_H^2)\, v\, w| &\leq \int_\Omega |f(u_H^2)\, v\, w| \\ &\leq c \int_\Omega (1 + (u_H)^{2q}) |v\, w| \qquad (\text{avec } 0 \leq q < 2\ ) \\ &\leq c\ \|v\|_{L^2(\Omega)}\ \|w\|_{L^2(\Omega)} \qquad (\text{en utilisant que (50)}) \\ &\leq c\ \|v\|_X\ \|w\|_X. \end{aligned} \tag{78}$$

Alors, en regroupant (77) et (78) on trouve

$$\forall\, v,\, w \in X \quad |g(v,w)| \;\leq\; c\|v\|_X\|w\|_X. \tag{79}$$
$$\tag{80}$$

3.1 Etude du problème 2

Théorème 3.1 *Si V vérifie l'hypothèse (7), si F vérifie les hypothèses (8)-(11) et si la forme bilinéaire g définie en (74) vérifie l'hypothèse de coercivité (75) alors il existe $\tilde{c} \in \mathbb{R}_+$ et $\delta_0 \in \mathbb{R}_+$ tels que pour tout $0 < h \leq \delta_0$ et $0 < H \leq \delta_0$ on ait :*

$$\|u - \tilde{u}_h^H\|_X \leq \tilde{c}[h + H^2]\|u\|_{H^2(\Omega)} \tag{81}$$

Ceci est semblable à l'estimation (21) dès que l'on choisit $h \sim H^2$.

Démonstration Dans un premier temps on montrera que le problème 2 admet une unique solution notée $\tilde{u}_h^H$ dans X_h et que celle-ci verifie la propriété suivante :

$$\|u - \tilde{u}_h^H\|_X \underset{h\to 0}{\longrightarrow} 0. \tag{82}$$

Considérons le problème continu suivant :
Trouver $\tilde{u}_*^H \in X$ telle que

$$\forall\, v \in X \qquad a(\tilde{u}_*^H, v) + \int_\Omega f(u_H^2)\tilde{u}_*^H v = \lambda_H \int_\Omega u_H v \tag{83}$$

et son approximation de Galerkin dans X_δ

$$\forall\, v_\delta \in X_\delta, \qquad a(\tilde{u}_\delta^H, v_\delta) + \int_\Omega f(u_H^2)\tilde{u}_\delta^H v_\delta = \lambda_H \int_\Omega u_H v_\delta, \tag{84}$$

avec $\delta = H$ ou h.
La forme bilinéaire g étant continue, coercive et symétrique sur X, le Thèorème de Lax-Milgram nous dit qu'il existe une unique solution au problème (83) dans X et une unique solution au problème (84) dans X_δ.
Montrons dans un premier temps que $\tilde{u}_*^H \in H^2(\Omega)$. Remarquons que le problème (83) est équivalent au problème suivant :
Trouver $\tilde{u}_*^H \in X$ telle que

$$-\Delta\tilde{u}_*^H + V\tilde{u}_*^H + f(u_H^2)\tilde{u}_*^H = \lambda_H u_H \tag{85}$$

Ainsi en remarquant que $f(u_H^2) \in L^2(\Omega)$ et en utilsant que $V \in L^p(\Omega)$, il vient $-\Delta\tilde{u}_H^* \in L^2(\Omega)$ et donc que $\tilde{u}_*^H \in H^2(\Omega)$ et

$$\begin{aligned} \|\tilde{u}_*^H\|_{H^2} &\leq c\,\|u_H\|_{L^2} \\ &\leq c\,\|u_H - u\|_{L^2} + \|u\|_{L^2} \\ &\leq c(1+H)\|u\|_{H^2}. \end{aligned} \tag{86}$$

Montrons maintenant que

$$\|\tilde{u}_*^H - \tilde{u}_\delta^H\|_X \underset{\delta\to 0}{\longrightarrow} 0. \tag{87}$$

Pour cela on va utiliser la coercivité de g dans X et le fait que $g(\tilde{u}_*^H - \tilde{u}_\delta^H, w_\delta) = 0$. Soit ϵ, la constante d'ellipticité de g, il découle de (83) et (84), que pour tout $w_\delta \in X_\delta$

$$\begin{aligned}
\xi\|\tilde{u}_*^H - \tilde{u}_\delta^H\|_X^2 &\leq g(\tilde{u}_*^H - \tilde{u}_\delta^H, \tilde{u}_*^H - \tilde{u}_\delta^H) \\
&\leq g(\tilde{u}_*^H - \tilde{u}_\delta^H, \tilde{u}_*^H - w_\delta) + g(\tilde{u}_*^H - \tilde{u}_\delta^H, w_\delta - \tilde{u}_\delta^H) \\
&\leq g(\tilde{u}_*^H - \tilde{u}_\delta^H, \tilde{u}_*^H - w_\delta) \\
&\leq c\|\tilde{u}_*^H - \tilde{u}_\delta^H\|_X \|\tilde{u}_*^H - w_\delta\|_X.
\end{aligned}$$

Ainsi en choisissant $w_\delta \in X_\delta$ telle que

$$\|\tilde{u}_*^H - w_\delta\|_X = \inf_{v_\delta \in X_\delta} \|\tilde{u}_*^H - v_\delta\|_X \leq c\,\delta \|\tilde{u}_*^H\|_{H^2(\Omega)}$$

on obtient

$$\|\tilde{u}_*^H - \tilde{u}_\delta^H\|_X \leq \frac{c}{\xi}\|\tilde{u}_*^H - w_\delta\|_X \leq \frac{c}{\xi}\,\delta\|\tilde{u}_*^H\|_{H^2(\Omega)} \underset{\delta\to 0}{\longrightarrow} 0. \tag{88}$$

De plus, on remarque u_H est la solution de (84) pour $\delta = H$, $\tilde{u}_h^H$ est la solution de (84) pour $\delta = h$ et

$$\|u - \tilde{u}_h^H\|_X \leq \|u - u_H\|_X + \|u_H - \tilde{u}_*^H\|_X + \|\tilde{u}_*^H - \tilde{u}_h^H\|_X.$$

Alors en utilisant (21) avec $\delta = H$ pour le terme $\|u - u_H\|_X$ et (87) avec $\delta = H$ pour le terme $\|u_H - \tilde{u}_*^H\|_X$ ainsi que (87) avec $\delta = h$ pour le terme $\|\tilde{u}_*^H - \tilde{u}_h^H\|_X$ dans l'inégalité précédente, on obtient (82).

Revenons à la démonstration de l'estimation (81). Soit $x_h \in X_h$, on note que

$$\|u - \tilde{u}_h^H\|_X \leq \|u - x_h\|_X + \|x_h - \tilde{u}_h^H\|_X, \tag{89}$$

ainsi d'après la coercivité de la forme bilinéaire b définie par (24) sur X, il existe une constante positive M telle que

$$\begin{aligned}
M\|x_h - \tilde{u}_h^H\|_X^2 \leq& b(x_h - \tilde{u}_h^H, x_h - \tilde{u}_h^H) \\
\leq& b(x_h - u, x_h - \tilde{u}_h^H) + b(u - \tilde{u}_h^H, x_h - \tilde{u}_h^H).
\end{aligned} \tag{90}$$

Notons $v_h = x_h - \tilde{u}_h^H$; le premier terme dans (90) sera traité comme suit

$$\begin{aligned}
b(u-\tilde{u}_h^H, v_h) =& \langle (A_u - \lambda)(u-\tilde{u}_h^H), v_h\rangle_{X',X} \\
=& -a(\tilde{u}_h^H, v_h) - \int_\Omega [f(u^2) - f(u_H^2)]\tilde{u}_h^H v_h - \int_\Omega f(u_H^2)\tilde{u}_h^H v_h \\
& + 2\int_\Omega f'(u^2)u(u-\tilde{u}_h^H)v_h + \lambda\int_\Omega u_h^H v_h \\
=& -\int_\Omega [f(u^2) - f(u_H^2)]\tilde{u}_h^H v_h + \lambda\int_\Omega u_h^H v_h \\
& + 2\int_\Omega f'(u^2)u(u-\tilde{u}_{Hh})v_h - \lambda_H\int_\Omega u_H v_h \\
=& -\int_\Omega [f(u^2) - f(u_H^2)]\tilde{u}_h^H v_h + 2\int_\Omega f'(u^2)u(u-\tilde{u}_h^H)v_h \\
& + (\lambda-\lambda_H)\int_\Omega u_h^H v_h - \lambda_H\int_\Omega (\tilde{u}_h^H - u)v_h + \lambda_H\int_\Omega (u-u_H)v_h \\
=& -\int_\Omega (u_H - u)\tilde{u}_h^H v_h w^H + 2\int_\Omega f'(u^2)u(u-\tilde{u}_h^H)v_h \\
& + (\lambda-\lambda_H)\int_\Omega u_h^H v_h - \lambda_H\int_\Omega (\tilde{u}_h^H - u)v_h + \lambda_H\int_\Omega (u-u_H)v_h
\end{aligned}$$

où $w^H = \frac{[f(u_H^2)-f(u^2)]}{u_H^2-u^2}(u+u_H)$. Ainsi, en utilsant (51) on a

$$|b(u-\tilde{u}_h^H, v_h)| \le \tilde{c}[|\lambda-\lambda_H| + \|u-u_H\|_{L^2(\Omega)} + \|u-\tilde{u}_h^H\|_{L^2(\Omega)}]\|v_h\|_{L^2(\Omega)}. \tag{91}$$

Il nous faut évaluer $\|u-\tilde{u}_h^H\|_{L^2(\Omega)}$, pour cela on utilisera l'inégalité suivante

$$\|u-\tilde{u}_h^H\|_{L^2} \le \|u-u_H\|_{L^2} + \|u_H - \tilde{u}_*^H\|_{L^2} + \|\tilde{u}_*^H - \tilde{u}_h^H\|_{L^2}. \tag{92}$$

Ainsi en utilisant un raisonnement de type Aubin pour évaluer les termes $\|u_H - \tilde{u}_*^H\|_{L^2}$ et $\|\tilde{u}_*^H - \tilde{u}_h^H\|_{L^2}$, on montrera que

$$\|u-\tilde{u}_h^H\|_{L^2} \le c\,H^2\|u\|_{L^2}. \tag{93}$$

En effet, on a

$$\|\tilde{u}_*^H - \tilde{u}_\delta^H\|_{L^2} = \sup_{w\in L^2(\Omega)} \frac{\displaystyle\int_\Omega w(\tilde{u}_*^H - \tilde{u}_\delta^H)}{\|w\|_{L^2}} \qquad \text{avec } \delta = H \text{ ou } h. \tag{94}$$

Notons p_w la solution du problème suivant : trouver $p_w \in X$ tel que

$$g(p_w, w) = \int_\Omega w(\tilde{u}_*^H - \tilde{u}_\delta^H) \qquad \forall w \in X.$$

En utilisant (75) et (79), le théorème de Lax-Milgram assure l'unicité de cette fonction p_w. De plus en remarquant que p_w est également solution du problème suivant : trouver $p_w \in X$ tel que

$$-\Delta p_w + V p_w + f(u_H^2)p_w = w,$$

on a les résultats de régularité suivant :

$$\begin{aligned} p_w &\in H^2(\Omega) && (95)\\ \|p_w\|_{H^2} &\leq c\|w\|_{L^2}. && (96)\end{aligned}$$

Ainsi on peut réécrire (94) de la façon suivante

$$\|\tilde{u}_*^H - \tilde{u}_\delta^H\|_{L^2} = \sup_{w\in L^2(\Omega)} \frac{g(p_w, \tilde{u}_*^H - \tilde{u}_\delta^H)}{\|w\|_{L^2}}. \quad (97)$$

Notons $p_{\delta,w} \in X_\delta$ (avec $\delta = H$ ou h), tel que

$$\|p_w - p_{\delta,w}\|_{H^1} \leq c\,\delta\|p_w\|_{H^2}. \quad (98)$$

Alors en utilisant le fait que pour tout $v_\delta \in X_\delta$, $g(v_h, \tilde{u}_*^H - \tilde{u}_\delta^H) = 0$ et (79), il vient que

$$\begin{aligned} \|\tilde{u}_*^H - \tilde{u}_\delta^H\|_{L^2} &\leq \sup_{w\in L^2(\Omega)} \frac{\|p_w - p_{\delta,w}\|_{H^1}\|\tilde{u}_*^H - \tilde{u}_\delta^H\|_{H^1}}{\|w\|_{L^2}}\\ &\leq c\ \delta\|\tilde{u}_*^H - \tilde{u}_\delta^H\|_{H^1}. && (99)\end{aligned}$$

Enfin en utilisant (86) et (88), il vient

$$\|\tilde{u}_*^H - \tilde{u}_\delta^H\|_{L^2} \leq c\,\delta^2\|\tilde{u}\|_{H^2} \quad (100)$$

et donc

$$\|u - \tilde{u}_h^H\|_{L^2} \leq c\,H^2\|u\|_{H^2}. \quad (101)$$

Revenons à l'estimation (90), alors

$$\begin{aligned} M\|x_h - \tilde{u}_h^H\|_X^2 \leq& b(x_h - u, x_h - \tilde{u}_h^H) + b(u - \tilde{u}_h^H, x_h - \tilde{u}_h^H)\\ \leq& b(x_h - u, x_h - \tilde{u}_h^H) + \tilde{c}[H^2\|u\|_{H^2(\Omega)} + h\|u - \tilde{u}_h^H\|_X]\|x_h - \tilde{u}_h^H\|_{L^2(\Omega)}\\ \leq& \tilde{c}\|x_h - \tilde{u}_h^H\|_X[\|u - x_h\|_X + H^2\|u\|_{H^2(\Omega)} + h\|u - \tilde{u}_h^H\|_X].\end{aligned}$$

Finalement, à partir de (89) on trouve que

$$\|u - \tilde{u}_h^H\|_X \leq \tilde{c}_1\|u - x_h\|_X + \tilde{c}_2H^2\|u\|_{H^2(\Omega)} \leq \tilde{c}[h + H^2]\|u\|_{H^2(\Omega)}. \quad (102)$$

Ce qui termine la preuve du Théorème 3.1. □

3.2 Etude du problème 3

Ici, on supposera également que la forme bilinéaire a vérifie la propriété suivante :

$$a(v, v) \geq \xi\|v\|_X^2 \quad \forall\, v \in X. \quad (103)$$

Théorème 3.2 *Si V vérifie l'hypothèse (7), si F vérifie les hypothèses (8)-(9) et (11), et si la forme bilinéaire a vérifie l'hypothèse de coercivité (103). Alors il existe $\overline{c} \in \mathbb{R}_+$ et $\delta_0 \in \mathbb{R}_+$ tels que pour tout $0 < h \leq \delta_0$ et $0 < H \leq \delta_0$ on ait*

$$\|u - \overline{u}_h^H\|_X \leq \overline{c}[h + H^2]\|u\|_{H^2(\Omega)} \tag{104}$$

Ceci est semblable à l'estimation (21) dès que l'on choisit $h \sim H^2$.

Démonstration Dans un premier temps on montrera que le problème 3 admet une unique solution notée $\overline{u}_h^H$ dans X_h et que celle-ci verifie la propriété suivante :

$$\|u - \overline{u}_h^H\|_X \underset{h\to 0}{\longrightarrow} 0. \tag{105}$$

Pour cela, on considère le problème suivant : Trouver $\overline{u}_*^H \in X$ telle que

$$\forall\, v \in X, \qquad a(\overline{u}_*^H, v) = \lambda_H \int_\Omega u_H v - \int_\Omega f(u_H^2)v \tag{106}$$

et son approximation de Galerkin dans X_δ : Trouver $\overline{u}_\delta^H \in X_\delta$ telle que

$$\forall\, v_\delta, \in X_\delta, \qquad a(\overline{u}_\delta^H, v_\delta,) = \lambda_H \int_\Omega u_H v_\delta, - \int_\Omega f(u_H^2)v_\delta, \tag{107}$$

avec $\delta = H$ ou h.
En remarquant que $f(u_H^2) \in L^2(\Omega)$ et en utilisant les propriétés de coercivité et de continuité de la forme bilinéaire a, on peut appliquer le Théorème de Lax-Milgram pour montrer que le problème (106) admet une unique solution dans X. De la même façon on obtient l'unicité de la solution du problème (107) dans X_δ.
Montrons maintenant que

$$\|\overline{u}_*^H - \overline{u}_\delta^H\|_X \underset{\delta\to 0}{\longrightarrow} 0. \tag{108}$$

Pour cela on va utiliser la coercivité de la forme bilinéaire a, en effet on a

$$\begin{aligned}
\xi\|\overline{u}_*^H - \overline{u}_\delta^H\|_X^2 &\leq a(\overline{u}_*^H - \overline{u}_\delta^H, \overline{u}_*^H - \overline{u}_\delta^H) \\
&\leq a(\overline{u}_*^H - \overline{u}_\delta^H, \overline{u}_*^H - w_\delta) + a(\overline{u}_*^H - \overline{u}_\delta^H, w_\delta - \overline{u}_\delta^H) \qquad \forall\, w_\delta \in X_\delta \\
&\leq a(\overline{u}_*^H - \overline{u}_\delta^H, \overline{u}_*^H - w_\delta) \\
&\qquad (\text{ en utilisant que } a(\overline{u}_*^H - \overline{u}_\delta^H, w_\delta) = 0,\ \forall\, w_\delta \in X_\delta) \\
&\leq c\|\overline{u}_*^H - \overline{u}_\delta^H\|_X \|\overline{u}_*^H - w_\delta\|_X.
\end{aligned}$$

Ainsi en choisissant w_δ telle que $\|\overline{u}_*^H - w_\delta\|_X = \inf_{v_\delta \in X_\delta} \|\overline{u}_*^H - v_\delta\|_X \leq c\delta \|\overline{u}_*^H\|_H^2(\Omega)$, on obtient (108).
De plus, on remarque que u_H est la solution de (107) pour $\delta = H$ et $\overline{u}_h^H$ est la solution de (107) dans $\delta = h$.
Alors on utilisant

$$\|u - \overline{u}_h^H\|_X \leq \|u - u_H\|_X + \|u_H - \overline{u}_*^H\|_X + \|\overline{u}_*^H - \overline{u}_h^H\|_X$$

et (21) avec $\delta = H$ pour le terme $\|u - u_H\|_X$ et (108) avec $\delta = H$ pour le terme $\|u_H - \overline{u}_*^H\|_X$ ainsi que (108) avec $\delta = h$ pour le terme $\|\overline{u}_*^H - u_h^H\|_X$ dans l'inégalité précédente, on obtient (105). Revenons, maintenant à la démonstration de l'estimation (104). Soit $x_h \in X_h$, on note que

$$\|u - \overline{u}_h^H\|_X \leq \|u - x_h\|_X + \|x_h - \overline{u}_h^H\|_X, \tag{109}$$

or d'après (103), il existe une constante positive ξ tel que

$$\begin{aligned} \xi\|x_h - \overline{u}_h^H\|_X^2 \leq & a(x_h - \overline{u}_h^H, x_h - \overline{u}_h^H) \\ \leq & a(x_h - u, x_h - \overline{u}_h^H) + a(u - \overline{u}_h^H, x_h - \overline{u}_h^H). \end{aligned} \tag{110}$$

Notons $w_h = x_h - \overline{u}_h^H$, le premier terme de l'inégalité précédente sera traité de la manière suivante

$$\begin{aligned} a(u - \overline{u}_h^H, w_h) =& \lambda \int_\Omega u w_h - \lambda_H \int_\Omega u_H w_h + \int_\Omega (f(u_H^2)u_H - f(u^2)u)w_h \\ =& (\lambda_H - \lambda) \int_\Omega u w_h + \lambda_H \int_\Omega (u - u_H) w_h \\ & + \int_\Omega (f(u_H^2) - f(u^2)) u_H w_h - \int_\Omega f(u^2)(u - u_H) w_h \\ =& (\lambda_H - \lambda) \int_\Omega u w_h + \lambda_H \int_\Omega (u - u_H) w_h \\ & + \int_\Omega (u - u_H) w^H u_H w_h - \int_\Omega f(u^2)(u - u_H) w_h \end{aligned}$$

où $w^H = \frac{[f(u_H^2) - f(u^2)]}{u_H^2 - u^2}(u + u_H)$. Ainsi, à partir de (8) et de (51), on obtient

$$|a(u - \overline{u}_h^H, w_h)| \leq \overline{c}\left[|\lambda - \lambda_H| + \|u - u_H\|_{L^2(\Omega)}\right] \|w_h\|_{L^2(\Omega)}. \tag{111}$$

En utilisant ceci dans (110), on trouve que

$$\xi\|x_h - \overline{u}_h^H\|_X^2 \leq a(x_h - u, x_h - \overline{u}_h^H) + \overline{c}[|\lambda - \lambda_H| + \|u - u_H\|_{L^2(\Omega)}]\|x_h - \overline{u}_h^H\|_{L^2(\Omega)}.$$

Par conséquent, d'après (77)

$$\xi\|x_h - \overline{u}_h^H\|_X^2 \leq \overline{c}\|x_h - \overline{u}_h^H\|_X[\|x_h - u\|_X + |\lambda - \lambda_H| + \|u - u_H\|_{L^2(\Omega)}] \tag{112}$$

alors

$$\|x_h - \overline{u}_h^H\|_X \leq \overline{c}[\|x_h - u\|_X + |\lambda - \lambda_H| + \|u - u_H\|_{L^2(\Omega)}]$$

et donc en utilisant (22) et (23) avec $\delta = H$, on obtient

$$\|x_h - \overline{u}_h^H\|_X^2 \leq \overline{c}[\|x_h - u\|_X + H^2\|u\|_{H^2(\Omega)}].$$

Finalement en revenant à (109) et en choisissant $x_h = \Pi_h u$ on retrouve que $\|u - \overline{u}_h^H\|_X \leq \overline{[}h + H^2]\|u\|_{H^2(\Omega)}$.

Ce qui termine la preuve du Théorème (3.2). □

4 Résultats Numériques

Dans cette section on s'intéresse aux problèmes discrets aux valeurs propres non linéaire associé à la fonctionnelle d'énergie suivante :

$$E(v) = \frac{1}{2}\int_\Omega \nabla v^2 + \frac{1}{2}\int_\Omega F(v^2)$$

avec $F(t^2) = t^{p+1}$, $2 < p \leq 3$, en dimension 2 sur un domaine carré $[0, \pi]^2$.

On a choisit de tester notre méthode sur plusieurs maillages grossiers de taille H_i, et un maillage fin h. À partir d'une triangulation $\mathcal{T}_0$, on construit les triangulations $\mathcal{T}_{n,1\leq n\leq 4}$ en découpant chaque triangle K appartenant à $\mathcal{T}_{n-1}$ en quatre triangles de même diamètre H_{n_K} tel que $H_{n_K} = \frac{H_{(n-1)K}}{2}$ (voir fig 4). Le sous-espace X_{H_n} obtenu est environ quatre fois plus grand que $X_{H_{n-1}}$ et vérifie $X_{H_0} \subset X_{H_n}$.

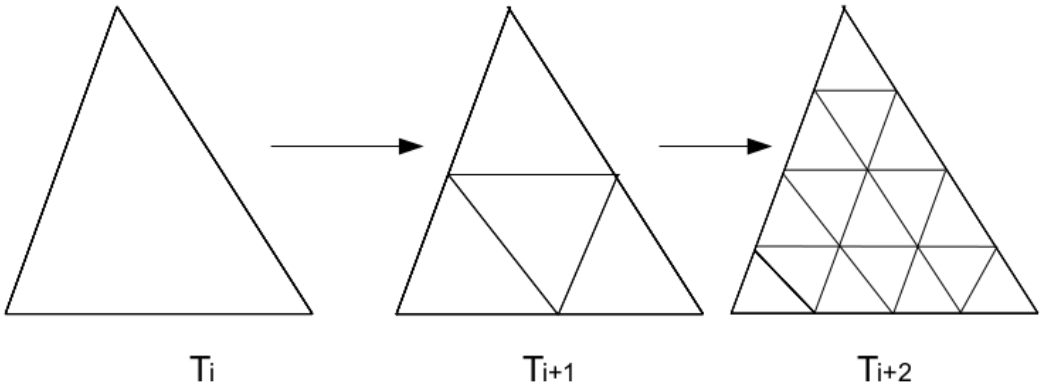

Fig. 1 – Construction de maillage emboités

On prendra $X_h = X_{H_4}$ et notera (u_h, λ_h) la solution du problème aux valeurs propres non linéaires sur X_h.

On remarque que les approximations de u calculées sur le maillage $\mathcal{T}_{H_0}$ sont aussi bonnes que u_h, bien que ce maillage soit 32 fois moins précis que le maillage $\mathcal{T}_{H_4}$. On remarque aussi que le temps de calcul a énormement diminué, en effet il est en moyenne 10 fois plus court, que celui nécessaire à résolution du problème non linéaire sur la grille fine. Les tables 1 et 2 illustrent les estimiations d'erreurs obtenues suivant les différentes méthodes utilisées et le raffinement du maillage grossier. De même la table 3 illustre les temps de calcul CPU obtenu,. Ces tests numériques ont été implémentés sous Freefem++ [22].

TAB. 1 – Comparaison des solutions obtenues par les méthodes à deux grilles pour p = 3

$$\|u - u_h\|_{H^1(\Omega_1)} = 0.00647426$$

T_n	Méthode à 2 Grilles			$\|u - u_{H_n}\|_{H^1(\Omega_1)}$
	$\|u - u_{H_n h}\|_{H^1(\Omega_1)}$	$\|u - \tilde{u}_{H_n h}\|_{H^1(\Omega_1)}$	$\|u - \overline{u}_{H_n h}\|_{H^1(\Omega_1)}$	
0	0.00647816	0.00658694	0.00660524	0.118264
1	0.00647449	0.00648174	0.00648297	0.0594255
2	0.00647426	0.00647474	0.00647482	0.0296258
3	0.00647426	0.00647428	0.00647429	0.0144717

TAB. 2 – Comparaison des valeurs propres obtenues par les méthodes à deux grilles pour $p = 3$.

$$|\lambda - \lambda_h| = 4.25 \times 10^{-5}$$

T_n	Méthode à 2 Grilles			$\|\lambda - \lambda_{H_n}\|$
	$\|\lambda - \lambda_{H_n h} - H^1(\Omega_1)$	$\|\lambda - \lambda_{H_n h}\|_{H^1(\Omega_1)}$	$\|\lambda - \overline{\lambda}_{H_n h}\|_{H^1(\Omega_1)}$	
0	3.46×10^{-5}	8.411×10^{-5}	9.69×10^{-5}	1.41×10^{-2}
1	4.05×10^{-5}	5.29×10^{-5}	5.61×10^{-5}	3.58×10^{-3}
2	4.20×10^{-5}	4.50×10^{-5}	4.59×10^{-5}	8.91×10^{-4}
3	4.24×10^{-5}	4.30×10^{-5}	4.33×10^{-5}	2.13×10^{-4}

TAB. 3 – Comparaison des temps de calcul CPU par les méthodes à deux grilles pour $p = 3$.

Résolution du problème non linéaire sur grille fine temps CPU = 121.46 s

T_n	Méthode à 2 Grilles		
	Problème 1	Problème 2	Problème 3
0	14.64 s	7.57 s	7.17 s
1	15.61 s	8.65 s	8.22 s
2	21.08 s	12.78 s	12.27 s
3	39.36 s	34.25 s	33.68 s

5 Annexe

Démonstration du Lemme 2.2 :

Montrons d'abord le résultat suivant : si $v \in X$, chaque terme de $E^H(v)$ est fini.
Soit $w \in X$, par définition de l'espace H^1, $\int_\Omega |\nabla w|^2 < +\infty$. Supposons d'abord $V \in L^p(\Omega)$ avec $p \geq 2$, et posons $p' = (1-p^{-1})^{-1}$. Ainsi $p' \leq 2$ et $v \in L^{p'}(\Omega)$, d'où $\int_\Omega Vv^2 \leq c\ \|V\|_{L^p}\|v\|^2_{L^{2p'}} \leq \infty$. Ainsi on obtient $E^H(v) \leq \infty$.
Montrons maintenant que $E^H(w)$ est bornée inférieurement sur l'ensemble $\{v \in X \mid \int_\Omega v^2 = 1\}$. Soit $v \in X$. En utilisant et la convexité de F, il vient que f est croissante et $f(u_H^2) > f(0)$. Ainsi, en raisonnant comme dans le Chapitre 1, on a

$$
\begin{aligned}
E^H(v) &= \int_\Omega (A\nabla v)\cdot\nabla v + \int_\Omega (V + f(u_H^2))v^2 \\
&\geq \alpha\|\nabla v\|^2_{L^2} - \|V\|_{L^p}\|v\|^2_{L^{2p'}} + (f(0))\|v\|^2_{L^2} \\
&\quad \text{(en utilisant (5))} \\
&\geq \alpha\|\nabla v\|^2_{L^2} - \|V\|_{L^p}\|v\|^{2-3/p}_{L^2}\|v\|^{3/p}_{L^6} + (f(0))\|v\|^2_{L^2} \\
&\geq \alpha\|\nabla v\|^2_{L^2} - C_6^{3/p}\|V\|_{L^p}\|v\|^{2-3/p}_{L^2}\|v\|^{3/p}_{H^1} + (f(0))\|v\|^2_{L^2} \\
&\geq \frac{\alpha}{2}\|\nabla v\|^2_{L^2} + \left(f(0) - \frac{3-2p}{2p}\left(\frac{3C_6^2\|V\|^{2p/3}_{L^p}}{p\alpha}\right)^{3/(2p-3)} - \frac{\alpha}{2}\right)\|v\|^2_{L^2}, \\
&\geq \left(f(0) - \frac{3-2p}{2p}\left(\frac{3C_6^2\|V\|^{2p/3}_{L^p}}{p\alpha}\right)^{3/(2p-3)} - \frac{\alpha}{2}\right)\|v\|^2_{L^2} \\
&\geq Z
\end{aligned}
$$

avec $Z = f(0) - \frac{3-2p}{2p}\left(\frac{3C_6^2\|V\|^{2p/3}_{L^p}}{p\alpha}\right)^{3/(2p-3)} - \frac{\alpha}{2}$. et C_6 est la constante venant de l'inégalité de Sobolev suivante : $\forall v \in X$, on a $\|v\|_{L^6} \leq C_6\|v\|_{H^1}$.
• Existence

Revenons maintenant à la preuve de l'existance d'un minimiseur de E^H :

1. Soit $(u_{*n}^H)_{n\in\mathbb{N}}$ une suite minimisante telle que

$$\begin{cases} u_{*n}^H \in X \\ \\ E^H(u_{*n}^H) \searrow_{n\to+\infty} I = \inf\{E^H(v),\, v \in X,\, \|v\|^2_{L^2(\Omega} = 1\} \\ \\ \|u_{*n}^H\|^2_{L^2(\Omega} = 1. \end{cases}$$

2. Montrons que cette suite est bornée dans X.
En remarquant que $\|u_{*n}^H\|_{L^2} = 1$, il vient que cette suite est bornée dans L^2. De plus en utilisant que pour tout $v \in X$ vérifiant $\displaystyle\int_\Omega v^2 = 1$, on a
$E^H(v) \geq Z$, on obtient $E^H(u_{*n}^H) \downarrow I \geq Z$. Ainsi pour n assez grand, $E^H(u_{*n}^H) \leq I + 1$ et

$$\int_\Omega |\nabla u_{*n}^H|^2 \leq I + 1 - Z$$

d'où

$$\|\nabla u_{*n}^H\|_{L^2} \leq \sqrt{I + 1 - Z}.$$

Donc la suite (u_{*n}^H) est bornée dans X.

3. On montre $u_{*n}^H \rightharpoonup u_*^H$ dans X.
La suite (u_{*n}^H) étant bornée dans X, on peut extraire une sous-suite $(u_{*n_k}^H)$ qui converge faiblement dans X vers $u_*^H \in X$.

L'ouvert Ω étant borné, les injections compactes de Sobolev nous disent $H^1(\Omega) \hookrightarrow L^2(\Omega)$. Ainsi on peut extraire de la suite $(u_{*n_k}^H)$ une suite, encore notée $(u_{*n_k}^H)$ pour simplifier, qui vérifie

$$u_{*n_k}^H \to \underline{u}_*^H \qquad \text{dans } L^2(\Omega)) \text{ fort.}$$

Et donc par unicité de la limite $u_*^H = \underline{u}_*^H$.

4. On montre

$$\begin{cases} (u_{*n}^H) \text{ bornée dans } X \\ u_{*n}^H \rightharpoonup u_*^H \text{ bornée dans } X \end{cases} \qquad \Longrightarrow \|u_*^H\|^2_{L^2(\Omega} = 1 \text{ et } \exists (u_{*n_k}^H)/E^H(u_*^H) \leq \liminf E^H(u_{*n_k}^H)$$

On utilsant que $\forall k \in \mathbb{N}$, $u_{n_k}^2 = \lambda$ et $u_{*n_k}^H \to \underline{u}_*^H$ dans $L^2(\Omega))$ fort, il vient immédiatement $\displaystyle\int_\Omega (u_*^H)^2 = 1$.

Les fonctionnelles $v \mapsto \displaystyle\int_\Omega V v^2$ et $v \mapsto \displaystyle\int_\Omega f(u_H^2)v^2$ étant continues pour la topologie forte dans X, il résulte qu'elles sont s.c.i. pour la topologie faible dans X.On en déduit que la fonctionnelle d'énergie E^H est s.c.i. pour la topologie faible dans X. Et donc

$$u_{*n_k}^H \rightharpoonup u_*^H \text{ dans } X \text{ faible, } E^H(u_*^H) \leq \lim_{n\leftarrow+\infty} E(u_{*n_k}^H) = I.$$

5. Il en résulte que u_*^H est solution du problème 33

• Unicité

Montrons que la fonctionnelle $\mathcal{E}^H$ est strictement convexe sur $\mathcal{K}$.
Commençons par regarder le terme $\rho \mapsto \nabla(\sqrt{\rho})^2$. Soient ρ_1 et $\rho_2 \in \mathcal{K}$ et $t \in [0,1]$, on pose $u_1 = \sqrt{\rho_1}$, $u_2 = \sqrt{\rho_2}$, $\rho = (1-t)\rho_1 + t\rho$ et $u = \sqrt{\rho}$. Pout tout $x \in \Omega$ on a

$$\begin{aligned} u(x)|\nabla\sqrt{\rho(x)}| =& u(x)|\nabla u(x)| = |\nabla(\frac{u^2(x)}{2})| = \frac{1}{2}|\nabla\rho(x)| \\ =& \frac{1}{2}|(1-t)\nabla\ \rho_1(x) + t\nabla\rho_2(x)| \\ =& |\sqrt{(1-t)}u_1(x)\nabla\ \sqrt{(1-t)}u_1(x) + \sqrt{t}u_2(x)\nabla\ \sqrt{t}u_2(x)| \\ \leq& [(1-t)u_1^2(c) + tu_2^2(x)]^{1/2}[(1-t)|\nabla u_1(x)|^2 + t|\nabla u_2(x)|^2]^{1/2} \\ =& u(x)[(1-t)|\nabla u_1(x)|^2 + t|\nabla u_2(x)|^2]^{1/2}. \end{aligned}$$

on obtient alors

$$|\nabla(\sqrt{\rho})^2| \leq (1-t)|\nabla u_1(x)|^2 + t|\nabla u_2(x)|^2$$

Donc $\rho \mapsto \nabla(\sqrt{\rho})^2$ est convexe sur $\mathcal{K}$, de même la fonctionnelle $\rho \mapsto V\rho$ est convexe sur $\mathcal{K}$. De plus la fonctionnelle $\rho \mapsto f(u_H^2)\rho$ est strictement convexe sur $\mathcal{K}$. Ceci montre que la fonctionnelle d'énergie $\mathcal{E}$ est strictement convexe sur $\mathcal{K}$.

Alors la fonctionnelle $\mathcal{E}$ admet un unique minimiseur $\rho^H = (u_*^H)^2$ sur $\mathcal{K}$, et donc u_*^H et $-u_*^H$sont les seuls minimiseurs de(33). On notera u_*^H la solution positive.

Deuxième partie

Schéma à deux grilles combinée à la méthode des bases réduites pour la résolution d' E.D.P paramétrées

Chapitre 4

Schéma à deux grilles combinée à la méthode des bases réduites pour la résolution d' E.D.P paramétrées

A two-grid finite-element/reduced basis scheme for the approximation of the solution of parameter dependent P.D.E

Rachida Chakir[1] and Yvon Maday[2]

[1,2] *UPMC Univ Paris 06, UMR 7598 LJLL, Paris, F-75005 France ; CNRS, UMR 7598 LJLL, Paris, F-75005 France.*

[2] *Division of Applied Mathematics, Brown University, Providence, RI, USA.*

Abstract

In the frame of optimization process in industrial framework, where numerical simulation is used at some stage, the same problem, modeled with partial differential equations depending on a parameter has to be solved many times for different sets of parameters. The reduced basis method may be successful in this frame and recent progress have permitted to make the computations reliable thanks to *a posteriori* estimators and to extend the method to non linear problems thanks to the "magic points" interpolation. However, it may not always be possible to use the code (for example of finite element type that allows for evaluating the elements of the reduced basis) to perform all the "off-line" computations required for an efficient performance of the reduced basis method. We propose here an alternating approach based on a coarse grid finite element the convergence of which is accelerated through the reduced basis and an improved post processing.

1 Introduction

Let X be a closed subspace of the Sobolev space $H^1(\Omega)$ over a bounded domain $\Omega \subset \mathbb{R}^d$ and $\mathcal{D}$ a set of parameter. We consider the following problem: given $\mu \in \mathcal{D}$, find $u(\mu) \in X$ such that

$$\forall\, v \in X, \quad a(u(\mu), v; \mu) = (f, v), \tag{1}$$

where $f \in L^2(\Omega)$ and a is a bilinear form, such that there exist $(\gamma_e(\mu), \gamma_c(\mu)) \in \mathbb{R} \times \mathbb{R}$, a couple of non negative constants that depends additionally on a parameter $\mu \in \mathcal{D}$ such that

$$a(v, v; \mu) \geq \gamma_e(\mu)\|v\|^2_{H^1(\Omega)} \qquad \forall\, v \in X \tag{2}$$

$$a(v, w; \mu) \leq \gamma_c(\mu)\|v\|_{H^1(\Omega)}\|w\|_{H^1(\Omega)} \qquad \forall\, (w, v) \in X. \tag{3}$$

In addition we assume that $u(\mu) \in H^2(\Omega)$ and there exist $c(\mu) > 0$ such that

$$\|u(\mu)\|_{H^2(\Omega)} \leq c(\mu)\|f\|_{L^2(\Omega)}. \tag{4}$$

In order to approximate the solution to this problem, one can use a standard numerical approach, as a finite element method, that provides a very accurate approximation of the solution $u(\mu)$ for any fixed value of the parameter μ. This accurate approximation will be called "truth approximation" in what follows. The computation of the truth approximation of $u(\mu)$ for many values of μ can however become very expensive as it has to be repeated for each parameter. The reduced basis method is an alternative that takes its roots upon the low "complexity" of the set of all solutions $\mathcal{M}^{\mathcal{D}} = \{u(\mu), \mu \in \mathcal{D}\}$ that can e.g. be measured by the Kolmogorov width [25] (see also [43]). This can for instance be formalized by the fact that for any $\epsilon > 0$, there exist a set of parameters $\mu_1, \mu_2, \cdots, \mu_N, \in \mathcal{D}$, where $N = N(\varepsilon)$ is reasonable, such that,

$$\forall \mu \in \mathcal{D} \quad \exists(\alpha_i(\mu)) \in \mathbb{R}^N, \qquad \|u(\mu) - \sum_{i=1}^{N} \alpha_i(\mu) u(\mu_i)\|_{H^1(\Omega)} \leq \varepsilon. \tag{5}$$

Based on the potential approximation property expressed above, the reduced basis method is in a Galerkin approach to the problem (1) for each new value of μ, within a space X_N spanned by N particular truth approximations of $u(\mu)$ corresponding to suitably chosen parameters μ.

To keep the interest of this method,

1. the parameters μ_i have to be adequately chosen,
2. the corresponding solution has to be properly calculated or approximated through an accurate discretization method (as a finite element method, for example)
3. the construction of the stiffness matrix $A(\mu)$ with entries $a(u(\mu_i), u(\mu_j); \mu)$ as to be done for each new value of μ.

All the expensive computations involving in the three previous steps are done off-line which allows to have online computations that scales only like powers of N and do not involve the dimension of the finite element space (see e.g. [34]).

Various recent contributions have permitted to extend the range of the reduced basis method, e.g. the *a posteriori* error estimates for validation and determination of the proper parameters μ_i's [44], the "magic points" , for generalization to nonlinear problems [18].

In an industrial framework, for optimization processes for instance these approaches have a great potential, unfortunately part of the off line computations require to enter in the code that computes the truth approximation which is not possible in case the simulation code has been bought or relies on a long evolution so that only a black box use of the code is possible. Those computations require indeed the use of some component involved in the implementation of discretization method which are not available to the user. As a consequence the reduced basis method cannot be efficiently implemented, an alternative needs to be proposed.

2 An alternating reduced basis method

Let us assume that the truth approximation is based on a $\mathbb{P}_1$-finite element code, capable of giving us a good enough approximation of the $u(\mu)$ in a finite element space X_h such that

$$\forall \mu \in \mathcal{D}, \qquad \|u(\mu) - u_h(\mu)\|_{H^1(\Omega)} \leq C(\mu) \inf_{v_h \in X_h} \|u(\mu) - v_h\|_{H^1(\Omega)},$$

with $C(\mu) = \dfrac{\gamma_c(\mu)}{\gamma_e(\mu)}$.

Moreover using that $\forall\, v \in H^2(\Omega)$, $\inf\limits_{v_h \in X_h} \|v - v_h\|_{H^1(\Omega)} \leq ch\|v\|_{H^2(\Omega)}$ and (4), we get

$$\forall \mu \in \mathcal{D}, \qquad ||u(\mu) - u_h(\mu)||_{H^1(\Omega)} \leq c_1(\mu)\ h\, \|f\|_{L^2(\Omega)} < Tol. \tag{6}$$

Where Tol is a a tolerance chosen in accordance to the final goal we have. In the standard reduced basis method we first compute the truth approximation $u_h(\mu_i)$, then form a discrete space

$$X_h^N = \text{Span}\{u_h(\mu_i), i = 1, .., N\}$$

and we build a Galerkin approximation of (1) in X_h^N: find $u_h^N(\mu) \in X_h^N$ such that,

$$\forall v \in X_h^N, \quad a(u_h^N(\mu), v; \mu) = (f, v). \tag{7}$$

In this case (5) is replaced by

$$\forall \mu \in \mathcal{D} \quad \exists(\alpha_i^h(\mu)) \in \mathbb{R}^N, \qquad \|u_h(\mu) - \sum_{i=1}^{N} \alpha_i^h(\mu) u_h(\mu_i)\|_{H^1(\Omega)} \leq \varepsilon. \tag{8}$$

Remark 2.1 *From (5), (6) and (8) we can see that for all $\mu \in \mathcal{D}$ there exist a set*
of $(\alpha_i^h(\mu))_{i=1:N} \in \mathbb{R}^N$ such that

$$\|u(\mu) - \sum_{i=1}^{N} \alpha_i^h(\mu) u_h(\mu_i)\|_{H^1(\Omega)} \leq \epsilon + c_2(\mu) h \|u(\mu)\|_{H^2(\Omega)},$$

where $c_2(\mu) = N\tilde{c}_2(\mu)$ and $\Pi_N u_h(\mu) = \sum\limits_{i=1}^{N} \alpha_i^h(\mu) u_h(\mu_i)$ is the best approximation of $u_h(\mu)$ in X_h^N.

The implantation of this reduced basis method involves the construction of the stiffness matrix $A_h(\mu)$ with entries $a(u_h(\mu_i), u_h(\mu_j); \mu)$.

In an industrial framework, the finite element code is often locked, so we can not decompose the construction of the stiffness matrix $A_h(\mu)$ into a series

of independent part that can be evaluated off line. This prevents us from employing the usual technique to compute quickly each stiffness matrix for a new value of μ, and take away the benefit of the reduced basis method (i.e. having a complexity depending only on N, independently of the dimension of the finite element space). First of all, let us remind that for a stable implementation of the reduced basis technique, it is required to build a better prepared basis than the one composed with the $u(\mu_i)$, usually a Gramm-Schmidt method is here advocated. We replace it here by the resolution of an eigenvalue problem: find $\xi \in X_h^N$ and $\lambda \in \mathbb{R}$ such that

$$\forall v \in X_h^N, \qquad \int_\Omega \nabla\xi\nabla v = \lambda \int_\Omega \xi v, \tag{9}$$

that provides $L^2(\Omega)$ and $H^1(\Omega)$ orthogonal eigenvectors $\xi_{i,BR}$ (chosen to be normalized in L^2). We note, that the $\xi_{i,BR}$ also constitute a second basis of the space X_h^N. Secondly we remark that, the standard reduced basis method aims at evaluating the coefficients intervening in the decomposition of $u_h^N(\mu)$ in the basis of the $\xi_{i,BR}$,

$$u_h^N(\mu) = \sum_{i=1}^{N} \beta_i^{BR} \xi_{i,BR}. \tag{10}$$

Those can appear as a substitute to the optimal coefficients

$$\beta_i^h(\mu) = \int_\Omega u_h(\mu)\xi_{i,BR}$$

of $\Pi_N u_h(\mu)$, the L^2-projection of $u_h(\mu)$ on X_h^N. This substitute is still good enough since, from Cea's Lemma we have

$$\begin{aligned} \|u(\mu) - u_h^N(\mu)\|_{H^1(\Omega)} &\leq c(\mu) \inf_{v\in X_h^N} \|u(\mu) - v\|_{H^1(\Omega)} \\ &\leq c(\mu)\ \|u(\mu) - \Pi_N u_h(\mu)\|_{H^1(\Omega)} \end{aligned}$$

then by using (5), (6) and (8) we derive

$$\|u(\mu) - u_h^N(\mu)\|_{H^1(\Omega)} \leq \epsilon + c_2(\mu)h\|u(\mu)\|_{H^2(\Omega} \tag{11}$$

Our alternative method first presented in [11] and illustrated by numerical results proving the potential interest of this alternative consists in proposing, another surrogate to $\beta_i^h(\mu)$ defined by

$$\beta_i^H(\mu) = \int_\Omega u_H(\mu)\xi_{i,BR}.$$

Since, the computation of $u_H(\mu)$, for $H >> h$ and $X_H \subset X_h$, is less expensive than the one of $u_h(\mu)$, the use of the industrial code with the parameter H to construct the $\beta_i^H(\mu)$is cheap enough. From this computation we derive

$$u_N^{Hh}(\mu) = \sum_{i=1}^{N} \beta_i^H(\mu)\xi_{i,BR}$$

in X_h^N. In what follow we explain in which case this can still be a very good approximation. We first notice that

$$\begin{aligned}\|u(\mu) - u_N^{Hh}(\mu)\|_{H^1(\Omega)} &\leq \|u(\mu) - \Pi_N u_h(\mu)\|_{H^1(\Omega)} + \|\Pi_N u_h(\mu) - u_N^{Hh}(\mu)\|_{H^1(\Omega)} \\ &\leq \epsilon + c_2(\mu)\ h\|u\|_{H^2(\Omega)} + \|\Pi_N u_h(\mu) - u_N^{Hh}(\mu)\|_{H^1(\Omega)}\end{aligned}$$

and we get that

$$\|\Pi_N u_h(\mu) - u_N^{Hh}(\mu)\|_{H^1(\Omega)} \leq \sum_{i=1}^{N} |\beta_i^h(\mu) - \beta_i^H(\mu)| \|\xi_{i,BR}\|_{H^1(\Omega)}.$$

Since the eigenvalue λ_i of the problem eigenproblem (8) are positives and can be ranked in increasing order $0 < \lambda_1 \leq \lambda_2 \leq ... < \lambda_N$, we can deduce from the L^2 orthogonality of the eigenfunctions that

$$\|\xi_{i,BR}\|_{H^1(\Omega)} \leq \sqrt{(1+\lambda_i)}\|\xi_{i,BR}\|_{L^2(\Omega)} \leq \sqrt{(1+\lambda_N)}.$$

Then

$$\|u_h^N(\mu) - u_N^{Hh}(\mu)\|_{H^1(\Omega)} \leq \sqrt{(1+\lambda_N)} \sum_{i=1}^{N} |\beta_i^h(\mu) - \beta_i^H(\mu)|.$$

Since $|\beta_i^h(\mu) - \beta_i^H(\mu)| \leq \|u_h(\mu) - u_H(\mu)\|_{0,\Omega}$ a classical Aubin-Nitsche argument[1] provides the following estimate:

$$\|u(\mu) - u_H(\mu)\|_{0,\Omega} \leq c(\mu) H \|u(\mu) - u_H(\mu)\|_{H^1(\Omega)} \leq cH^2 \|u(\mu)\|_{H^2(\Omega)}.$$

Therefore

$$\|u_h^N(\mu) - u_N^{Hh}(\mu)\|_{H^1(\Omega)} \leq N\sqrt{(1+\lambda_N)}[\bar{c}_1(\mu)H^2 + \bar{c}_2(\mu)h^2]\|u(\mu)\|_{H^2(\Omega)}.$$

By regrouping all this estimate we finally get that

$$\|u(\mu) - u_N^{Hh}(\mu)\|_{H^1(\Omega)} \leq \varepsilon + c_3(\mu)h + c_4(\mu)H^2$$

where $c_3(\mu) = c_3(N, \mu)$ and $c_4(\mu) = c_4(N, \mu)$ which is asymptotically similar to (11) when we choose $h \sim H^2$.

In the case of an higher order finite element approximation, $\mathbb{P}_k$, we can as well use an Aubin-Nitsche argument to get the improved error estimation. First we define $\Phi_{i,BR} \in X$, such that

$$\forall v \in X, \qquad a(v_h, \Phi_{i,BR}; \mu) = \int_\Omega \xi_{i,BR} v,$$

[1] Actually, the convergence results stated here either require that there is no corner or edge type singularities in the solutions — of the primal or dual problem for the Aubin-Nitsche argument — or that we relax somehow the definition of h and H being here a parameter associated with the grid size and the way the global refinement is done for convergence but not the size of the finer elements that should be defined such that the error bound by a conatant times h or H holds.

hence

$$\beta_i^h(\mu) - \beta_i^H(\mu) = \int_\Omega (u_h(\mu) - u_H(\mu))\xi_{i,BR} = a(u_h(\mu) - u_H(\mu), \Phi_{i;BR}; \mu) \quad (9).$$

Since $X_H \subset X_h$ we obtain, from the definition of $u_h(\mu)$ and $u_H(\mu)$,

$$\forall \chi_H \in X_H \qquad a(u_h(\mu) - u_H(\mu), \chi_H; \mu) = 0,$$

thus using this in (9) we derive

$$\forall \chi_H \in X_H, \quad \beta_i^h(\mu) - \beta_i^H(\mu) = a(u_h(\mu) - u_H(\mu), \Phi_{i,BR} - \chi_H; \mu).$$

Therefore by choosing χ_H such that

$$\|\Phi_{i,BR} - \chi_H\|_X = \inf_{v_H \in X_H} \|\Phi_{i,BR} - v_H\|_X$$

we have

$$|\beta_i^h(\mu) - \beta_i^H(\mu)| \leq c(\mu)\|u_h(\mu) - u_H(\mu)\|_{H^1(\Omega)}\|\Phi_{i,BR} - \chi_H; \mu)\|_{H^1(\Omega)} \leq c(\mu)H^{2k}$$

and then by proceeding as in the $\mathbb{P}_1$ approximation we obtain that

$$\|u(\mu) - u_N^{Hh}(\mu)\|_{H^1(\Omega)} \leq \varepsilon + c_5(\mu)h^k + c_6(\mu)H^{2k}$$

(where $c_5(\mu) = c_5(N, \mu)$ and $c_6(\mu) = c_6(N, \mu)$). Finally, we get to the same conclusion as previously by choosing $h \sim H^2$.

3 Post-processing

To improve even further the accuracy of the approach we propose to do a simple post processing of the results.
Let $\underline{\beta}^H(\mu_j)$ be the vector $(\beta_i^H(\mu_j))_{1 \leqslant i \leqslant N}$ and $\underline{\beta}^h(\mu_j)$ the one corresponding to the $(\beta_i^h(\mu_j))_{1 \leqslant i \leqslant N}$. We decide to improve the computation of the $\underline{\beta}^H(\mu)$, by a post-processing that will insure that for each parameters $\mu = (\mu_j)_{j=1,\cdots,N}$ that are used in the construction of the reduced basis, the method returns exactly $u_h(\mu_j)$.
Indeed contrarily to $u_h(\mu)$, that we do not want to compute for a large number of values of μ, the truth solutions $u_h(\mu_j), j = 1, \cdots, N$ have been actually computed. In order to define this post-processing, we consider the linear transformation $\mathcal{F} : \mathbb{R}^N \to \mathbb{R}^N$, that maps $\underline{\beta}^H(\mu_j)$ on to $\underline{\beta}^h(\mu_j)$. The post processing consist in applying it to all the vector $\underline{\beta}^H(\mu)$.

Let T be the matrix associated to the transformation $\mathcal{F}$. For large values of N, the solutions $u_h(\mu_j), j = 1, \cdots, N$ mays become almost linearly dependent which results in a bad conditioning of the matrix T. This loss of stability may result in an important deterioration of the vectors $\underline{\beta}^H(\mu)$. To avoid this problem we propose to map only the first solutions in the previous set and thus

construct an alternative matrix denoted T_k, $1 \leq k \leq N$ verifying:

$$(T_k)\left(\underline{\beta}^H(\mu_{p_1}), \cdots, \underline{\beta}^H(\mu_{p_k}), \underline{\gamma}_{k+1}, \cdots, \underline{\gamma}_N\right)^t = \left(\underline{\beta}^h(\mu_{p_1}), \cdots, \underline{\beta}^h(\mu_{p_k}), \underline{\gamma}_{k+1}, \cdots, \underline{\gamma}_N\right)^t$$

where the N vectors $\underline{\gamma}_k$ are constructed by a Gram - Schmidt method such that $\underline{\gamma}_1 = \dfrac{\underline{\beta}_H(\mu_{p_1})}{\|\underline{\beta}_H(\mu_{p_1})\|_2}$, and $\underline{\gamma}_k \in \text{Span}\{\underline{\beta}^H(\mu_{p_1}), \cdots, \underline{\beta}^H(\mu_{p_k})\}$.

The set $(\mu_{p_k})_{1\leq k\leq N}$ is identical to the one used in the construction of the reduced basis, but it has been arranged differently. Indeed for each iteration k, we choose μ_{p_k} among the $N-k$ parameters $(\mu_{p_q})_{,k\leq q\leq N}$, such that $\max\limits_{1\leq j\leq N} \|T_k\beta^H(\mu_j) - \beta^h(\mu_j)\|_\infty$ is the smallest. We notice that, at the end, the matrix T_N and T are similar.

We denote by $(\alpha^{k,H}(\mu))_{i=1,\cdots,N}$ the vector obtained by doing the following matrix-vector product $(T_k)(\underline{\beta}^H(\mu))$, with $1 \leq k \leq N$ chosen in such a way that the condition number of the matrix T_k is moderate enough. Finally the solution u_N^{Hh} is replaced by $\sum\limits_{i=1}^{N} \alpha_i^{k,H}(\mu)\xi_{i,BR}$.

4 Numerical results

The problems we consider in this section are in 2 dimensions. From an original coarse triangulation $\mathcal{T}_{H_0}$, we built successive refined triangulations $\mathcal{T}_{H_i, 1\leq i\leq 4}$ by recursively splitting each triangle K in $\mathcal{T}_{H_{i-1}}$ into four triangles with equal diameter H_{i_K} such that $H_{i_K} = \dfrac{H_{(i-1)_K}}{2}$. We get a superspace X_{H_i} about four times larger than $X_{H_{i-1}}$ that satisfies $X_{H_0} \subset X_{H_n}$.

The set of μ_i used to build the subspace X_h^N is found using a proper orthogonal decomposition (POD) technique :
Step 1: Solve a $\mathbb{P}_k$ finite element approximation of the problem (1) for n=100 different values of μ and obtain a set of n "snapshots" $u_h(\mu_j)$, $j = 1, \cdots, n$.
Step 2: Compute the stiffness matrix S defined by

$$(S_{jk})_{1\leq j,k\leq n} = \int_\Omega \nabla(u_h(\mu_j))\nabla(u_h(\mu_k))$$

and the mass matrix M defined by

$$(M_{jk})_{1\leq j,k\leq n} = \int_\Omega u_h(\mu_j)u_h(\mu_k).$$

Step 3: Solve the eigenvalue problem $SW = \lambda MW$.
Step 4: Choose the N eigenvectors $(W_j)_{1\leq j\leq N}$ associated to the N largest eigenvalues and then rewrite them as a linear combination of the snapshots i.e

$$W_j = \sum_{k=1}^{N} \gamma_k u_h(\mu_k)$$

Step 5 : For each of the N eigenvectors W_i, previously chosen, we take $u_h(\mu_i)$, among the snapshots, with $i = \underset{1\leq j\leq n}{\text{argmax}}(\gamma_j)$ to generate our reduced basis X_h^N.

We denote by $u_N^{hP}(\mu)$, the H^1 projection of $u_h(\mu)$ on the basis of the $\xi_{i,BR}$, defined by

$$u_N^{hP}(\mu) = \sum_{i=1}^{N} \beta_i^h(\mu)\xi_{i,BR}.$$

It is the best we can expect from the reduced basis, that is one of the ingredient entering in the approximation.

4.1 Example 1

We first consider the nonlinear problem: find $u \in H^1(\Omega)$ such that

$$\begin{array}{cc} -\Delta u + u^3 = \sin(x)\sin(y) & \text{in } \Omega = [0,1]^2 \setminus (]\frac{1}{2},1[^2) \text{ (L-shape domain)} \\ -\alpha u + \frac{\partial u}{\partial n} = y(1-y) & \text{on } \Gamma_F = \{(1,y), y \in [0,\frac{1}{2}]\} \\ u = \eta\ xy(1-y)(1-x) & \text{on } \Gamma_D = \partial\Omega \setminus \Gamma_F \end{array}$$

In this example, the set of parameters, $\mu = (\alpha, \eta)$, that we use is varying in $[1,37] \times [1,100]$. Let be $\mu_{H_i} = \underset{\mu=(\beta,\eta)\in[1,37]\times[1,100]}{\text{argmax}} \{||u(\mu) - u_N^{hH_i}(\mu)||_{1,\Omega}\}$ and $\mu_h = \underset{\mu=(\beta,\eta)\in[1,37]\times[1,100]}{\text{argmax}} \{||u(\mu) - u^h(\mu)||_{1,\Omega}\}$.

The results of the $\mathbb{P}_1$ approximation are showed in the table 1. We first remark that we need at least $N = 10$ elements in the reduced basis to recover the truth error. Second, before post-processing we note that the H_1-error made with the solution $u_N^{hH_2}$ is close to the one made with u_h, for any value of μ in $\mathcal{D}$, despite the fact that $\mathcal{T}_{H_2}$ is eight times less accurate than $\mathcal{T}_{H_4}$, at least for $N \geq 10$. We also note a small deterioration of the evaluation of the solution, when N rises, confirming that the constant $c_4(N)$ is growing with N. Finally, we note that the post-processing improved even more the approximation since it allows to recover the truth error even starting from the computations of the coarsest solution $u_N^{hH_0}$, at least if we use the proper number of reduced elements (10 or 15) , which is a very substantial savings. We note also that the reduction of indices in the post-processing is used, even it is important since, in the case $N = 15$ the error $\|u(\mu_{H_1}) - u_N^{hH_1}(\mu_{H_1})\|_{1,\Omega}$ with the full matrix is 0.50.

Table 1: Error for the example 1 with $X_h = \{v \in \mathcal{C}^0(\overline{\Omega}),\ v_{|T} \in \mathbb{P}_1(T),\ T \in \mathcal{T}_{H_4}\}$

$$||u(\mu_h) - u_h(\mu_h)||_{1,\Omega} = 3.3 \times 10^{-2}$$

N	k	i	$\|\|u(\mu_{H_i}) - u_N^{hH_i}(\mu_{H_i})\|\|_{1,\Omega}$ with post-processing	$\|\|u(\mu_{H_i}) - u_N^{hH_i}(\mu_{H_i})\|\|_{1,\Omega}$ without post-processing	$\|\|u(\mu_{H_i}) - u_{H_i}(\mu_{H_i})\|\|_{1,\Omega}$
5	$\|\|u(\mu_h) - u_h^{BR}(\mu_h)\|\|_{1,\Omega} = 0.19$				
	5	0	0.15	0.13	0.49
		1	0.17	0.16	0.28
		2	0.19	0.18	0.15
		3	0.19	0.19	7.3×10^{-2}
10	$\|\|u(\mu_h) - u_h^{BR}(\mu_h)\|\|_{1,\Omega} = 3.6 \times 10^{-2}$				
	10	0	3.6×10^{-2}	0.35	0.49
		1	3.5×10^{-2}	6.8×10^{-2}	0.28
		2	3.5×10^{-2}	3.8×10^{-2}	0.15
		3	3.6×10^{-2}	3.5×10^{-2}	7.3×10^{-2}
15	$\|\|u(\mu_h) - u_h^{BR}(\mu_h)\|\|_{1,\Omega} = 3.4 \times 10^{-2}$				
	15	0	3.5×10^{-2}	0.47	0.49
	7	1	3.7×10^{-2}	0.14	0.28
	15	2	3.4×10^{-2}	3.4×10^{-2}	0.15
		3	3.4×10^{-2}	3.4×10^{-2}	7.3×10^{-2}
20	$\|\|u(\mu_h) - u_h^{BR}(\mu_h)\|\|_{1,\Omega} = 3.3 \times 10^{-2}$				
	20	0	5.5×10^{-2}	0.56	0.49
		1	3.4×10^{-2}	0.20	0.28
		2	3.4×10^{-2}	4.8×10^{-2}	0.15
		3	3.4×10^{-2}	3.4×10^{-2}	7.3×10^{-2}

4.2 Example 2

The second problem is a convection dominated problem : find $u \in H^1(\Omega)$ such that

$$\begin{aligned} -(0.01)\Delta u + v \cdot \nabla u = 0 \quad & \text{in } \Omega = [0,1]^2 \\ u = x^2 \quad & \text{on } \Gamma_1 = \{(1,y), y \in [0,1]\} \\ u = y^2 \quad & \text{on } \Gamma_2 = \{(x,1), x \in [0,1]\} \\ u = 0 \quad & \text{on } \Gamma_3 = \partial\Omega \setminus (\Gamma_1 \cup \Gamma_2). \end{aligned}$$

where v is such as $v = (\cos\mu,\ \sin\mu)$. Here, the varying parameter is the angle of the convection $\mu \in [0, \frac{\pi}{2}]$. Let be $\mu_{H_i} = \underset{\mu\in[0,\frac{\pi}{2}]}{\text{argmax}}\{||u(\mu) - u_N^{hH_i}(\mu)||_{1,\Omega}\}$ and $\mu_h = \underset{\mu\in[0,\frac{\pi}{2}]}{\text{argmax}}\{||u(\mu) - u^h(\mu)||_{1,\Omega}\}$. The table 3 shows the results of $\mathbb{P}_1$ approximation while the table 2 shows the $\mathbb{P}_2$ ones. We can make the same conclusion than in the previous example : this combined method (reduced basis + two grids) is thus even improved by the trivial postprocessing. Note that the mathematical justification of this last ingredient is still missing.

Table 2: Error for the example 2 with $X_h = \{v \in \mathcal{C}^0(\overline{\Omega}),\ v_{|T} \in \mathbb{P}_2(T),\ T \in \mathcal{T}_{H_3}\}$

$$\|u(\mu_h) - u_h(\mu_h)\|_{1,\Omega} = 3.5 \times 10^{-3}$$

N	k	i	$\|\|u(\mu_{H_i}) - u_N^{hH_i}(\mu_{H_i})\|\|_{1,\Omega}$ with post-processing	$\|\|u(\mu_{H_i}) - u_N^{hH_i}(\mu_{H_i})\|\|_{1,\Omega}$ without post-processing	$\|\|u(\mu_{H_i}) - u_{H_i}(\mu_{H_i})\|\|_{1,\Omega}$
5	$\|\|u(\mu_h) - u_h^{BR}(\mu_h)\|\|_{1,\Omega} = 1.41 \times 10^{-2}$				
	5	0	6.4×10^{-2}	6.5×10^{-2}	0.11
		1	6.1×10^{-2}	6.1×10^{-2}	3.3×10^{-2}
		2	6.1×10^{-2}	6.1×10^{-2}	8.7×10^{-3}
10	$\|\|u(\mu_h) - u_h^{BR}(\mu_h)\|\|_{1,\Omega} = 3.5 \times 10^{-3}$				
	10	0	3.5×10^{-3}	4.1×10^{-2}	0.16
		1	3.5×10^{-3}	5.6×10^{-3}	5.1×10^{-2}
		2	3.5×10^{-3}	3.5×10^{-3}	1.4×10^{-2}
15	$\|\|u(\mu_h) - u_h^{BR}(\mu_h)\|\|_{1,\Omega} = 3.5 \times 10^{-3}$				
	15	0	3.5×10^{-3}	5.8×10^{-2}	0.16
		1	3.5×10^{-3}	7.8×10^{-3}	5.1×10^{-2}
		2	3.5×10^{-3}	3.5×10^{-3}	01.4×10^{-2}

Table 3: Error for the example 2 with $X_h = \{v \in \mathcal{C}^0(\overline{\Omega}),\ v_{|T} \in \mathbb{P}_1(T),\ T \in \mathcal{T}_{H_4}\}$

$$\|u(\mu_h) - u_h(\mu_h)\|_{1,\Omega} = 3.5 \times 10^{-2}$$

N	k	i	$\|\|u(\mu_{H_i}) - u_N^{hH_i}(\mu_{H_i})\|\|_{1,\Omega}$ with post-processing	$\|\|u(\mu_{H_i}) - u_N^{hH_i}(\mu_{H_i})\|\|_{1,\Omega}$ without post-processing	$\|\|u(\mu_{H_i}) - u_{H_i}(\mu_{H_i})\|\|_{1,\Omega}$
5	$\|\|u(\mu_h) - u_h^{BR}(\mu_h)\|\|_{1,\Omega} = 3.6 \times 10^{-2}$				
	5	0	5.7×10^{-2}	0.17	0.45
		1	6.2×10^{-2}	9.0×10^{-2}	0.24
		2	6.5×10^{-2}	7.0×10^{-2}	0.12
		3	6.6×10^{-2}	6.7×10^{-2}	6.1×10^{-2}
10	$\|\|u(\mu_h) - u_h^{BR}(\mu_h)\|\|_{1,\Omega} = 3.5 \times 10^{-2}$				
	10	0	3.5×10^{-2}	0.244	0.53
		1	3.5×10^{-2}	9.1×10^{-2}	0.31
		2	3.5×10^{-2}	4.2×10^{-2}	0.16
		3	3.5×10^{-2}	3.6×10^{-2}	7.9×10^{-2}
15	$\|\|u(\mu_h) - u_h^{BR}(\mu_h)\|\|_{1,\Omega} = 3.5 \times 10^{-2}$				
	15	0	3.5×10^{-2}	0.36	0.53
		1	3.5×10^{-2}	9.8×10^{-2}	0.31
		2	3.5×10^{-2}	4.2×10^{-2}	0.16
		3	3.5×10^{-2}	3.6×10^{-2}	7.9×10^{-2}
20	$\|\|u(\mu_h) - u_h^{BR}(\mu_h)\|\|_{1,\Omega} = 3.5 \times 10^{-2}$				
	13	0	3.5×10^{-2}	0.37	0.53
		1	3.5×10^{-2}	0.13	0.31
		2	3.5×10^{-2}	4.6×10^{-2}	0.16
		3	3.5×10^{-2}	3.6×10^{-2}	7.9×10^{-2}

List of Figures

Bibliographie

[1] H. Abboud and T. Sayah, *A full discretization of the time-dependent Navier-Stokes equations by a two-grid scheme*, M2AN Math. Model. Numer. Anal., Vol.42, N1, pp.141-174, (2008).

[2] I. Babuska and J. Osborn *Eigenvalue Problems*, In Handbook of Numerical Analysis, Vol. II, North-Holland: Amsterdam, 17–351(1991).

[3] X. Blanc and E. Cancès, *Nonlinear instability of density-independent orbital-free kinetic energy functionals*, J. Chem. Phys. 122 (2005) 214106.

[4] C. Bernadi, Y. Maday et F. Rapettti, *Discrétisation variationnelles de problèmes aux limites elliptiques* , Springer, (2000).

[5] Cancès, E., and Le Bris, C. , *Can we outperform the DIIS approach for electronic structure calculations?*. Int. J. Quantum Chem. 79, 82-90 (2000).

[6] E.Cancès, *SCF algorithms for Kohn-Sham models with fractional occupation numbers*. J. Chem. Phys. 114, 10616-10623 (2001).

[7] E. Cancès, R. Chakir and Y. Maday, *Numerical analysis of nonlinear eigenvalue problems*, Preprint arXiv:0905.1645.

[8] E. Cancès, R. Chakir and Y. Maday, *Numerical analysis of the planewave discretization of Kohn-Sham and related models*, in preparation.

[9] E. Cancès, M. Defranceschi, W. Kutzelnigg, C. Le Bris and Y. Maday, *Computational quantum chemistry: a primer*, in Handbook of numerical analysis, Volume X, pp 3–270, North-Holland,Amsterdam, (2003).

[10] C. Canuto, M.Y. Hussaini, A. Quarteroni and T.A. Zang ,*Spectral methods*, Springer (2007).

[11] R. Chakir and Y Maday, *Une méthode combinée d'éléments finis à deux grilles/bases réduites pour l'approximation des solutions d'une E.D.P. paramétrique*, C. R. Acad. Sci. Paris, Ser. I 347 435 - 440. (2009).

[12] X. Dai and A. Zhou *Three-scale finite element discretizations for quantum eigenvalue problems*, SIAM J. Numer. Anal., 46, 295-324 (2008).

[13] C. Dion and E. Cancès, *Spectral method for the time-dependent Gross-Pitaevskii equation with harmonic traps*. Phys. Rev. E. 67, 046706 (2003).

[14] RM Dreizler and EKU Gross, *Density functional theory: an approach to the quantum many-body problem*, Springer-Verlag Berlin (1990).

[15] A. Ern and J.-L. Guermond, *Theory and practice of finite elements*, Springer, (2004).

[16] E. Fermi *A statistical method for the determination of some atomic properties and the application of this method to the theory of the periodic system of elements* [a translation into english can be found in March, 1975]. Z. Phys., 48 :73–79, (1928).

[17] M.A. Grelp and A.T. Patera *Reduced-basis approximation for time-dependent parametrized partial differential equations*. M2AN 39(1),157-181(2005).

[18] M.A. Grepl, Y. Maday, N.C. Nguyen, and A.T. Patera, *Efficient reduced-basis treatment of nonaffine and nonlinear partial differential equations*, M2AN, 41(3):575-605, (2007).

[19] D. Gilbarg and Trudinger, *Elliptic partial differential equations of second order*, 3rd edition. Springer (1998).

[20] V. Girault, J.L. Lions, *Two-grid finite-element schemes for the steady Navier-Stokes problem in polyhedra*, Port. Math. (N.S.) 58, No.1, 25-57, 2001.

[21] D. Hartree *The wave mechanics of an atom with a non-Coulomb central field. part I. theory and methods*. Proc. Camb. Phil. Soc., 24 :89–312, (1928).

[22] F. Hecht and O. Pironneau, *Freefem++*, `http://www.freefem.org/ff++/`.

[23] P. Hohenberg and W. Kohn *Inhomogeneous electron gas*. Phys. Rev. A, 136(3) :B864– B871, (1964).

[24] W. Kohn and L.J. Sham. *Self-consistent equations including exchange and correlation effects*. Phys. Rev. A, 140(4) :A1133–A1138, (1965).

[25] Kolmogoroff, A., Über die beste Annäherung von Funktionen einer gegebenen Funktionenklasse *Anals of Math.* **37**, 107—110,(1963).

[26] M. Levy *Electron densities in search of Hamiltonians*. Phys. Rev. A, 26(3) :1200–1208, (1982).

[27] L. Landau et E. Lifchitz *Mécanique quantique*. Éditions MIR, Moscou (1966).

[28] B. Langwallner, C. Ortner and E. Süli *Existence and Convergence Results for the Galerkin Approximation of an Electronic Density Functional*. Submitted (2009).

[29] W. Layton and W. Lenferink, *Two-level Picard and modified Picard methods for the Navier-Stokes equations*, Appl. Math. Comp., 69, 263-274, (1995).

[30] E.H. Lieb and B. Simon. *On solutions to the Hartree-Fock problem for atoms and molecules*. J. Chem. Phys., 61 :735–736, (1974).

[31] E.H. Lieb, *Thomas-Fermi and related theories of atoms and molecules*, Rev. Mod. Phys. 53 603-641(1981).

[32] P.-L. Lions *Solutions of Hartree-Fock equations for Coulomb systems* Commun. Math. Phys, 109 :33–97 (1987).

[33] L. Machiels, Y. Maday, I. B. Oliveira, A. T. Patera, and D.V. Rovas *Output bounds for reduced-basis approximations of symmetric positive definite eigenvalue problems*. C. R. Acad. Sci. Paris, Série I, 331(2):153–158 (2000).

[34] Y. Maday, Yvon, *Reduced basis method for the rapid and reliable solution of partial differential equations*, in International Congress of Mathematicians. Vol. III, 1255–1270, Eur. Math. Soc., Zürich (2006).

[35] Y. Maday, N. C. Nguyen, A. T. Patera, and G. S. H. Pau. *A general multipurpose interpolation procedure: the magic points*. Comm. Pure. Appl. Anal., 8(1):383–404, January (2009).

[36] Y. Maday, A.T. Patera, and G. Turinici *Global a priori convergence theory for reduced-basis approximation of single parameter symmetric coercive elliptic partial differential equations*. C. R. Acad. Sci. Paris, Série I, 335(3):289–294 (2002).

[37] Y. Maday and G. Turinici, *Error bars and quadratically convergent methods for the numerical simulation of the Hartree-Fock equations*, Numer. Math. 94, 739-770(2003) .

[38] L.P. Pitaevskii and S. Stringari, *Bose-Einstein condensation*. Clarendon Press (2003).

[39] M. Marion and J. Xu, *Error estimates on a new nonlinear Galerkin method based on two-grid finite elements,* SIAM J. Numer. Anal., 32: 1170-1184, (1995).

[40] A. K. Noor and J. M. Peters *Reduced basis technique for nonlinear analysis of structures*, AIAA Journal, 18(4):455–462, (1980).

[41] R.G. Parr and W. Yang. *Density-Functionnal Theory of Atoms and Molecules*. Oxford Science Publications, (1989).

[42] J. S. Peterson *The reduced basis method for incompressible viscous flow calculations*. SIAM J. Sci. Stat. Comput., 10(4):777–786, (1989).

[43] Pinkus, A., *n-Widths in Approximation Theory*, Springer-Verlag, Berlin (1985).

[44] C Prud'homme, DV Rovas, K Veroy, L Machiels, Y Maday, AT Patera, and G Turinici, *Reliable real-time solution of parametrized partial differential equations: Reduced-basis output bound methods*, J Fluids Engineering, 124:70–80 (2002).

[45] P.A. Raviart et J.M. Thomas *Introduction à l'analyse numérique des équations aux dérivées partielles*, Masson, 3ème édition (1992).

[46] S. Sen, K. Veroy, D.B.P. Huynh, S Deparis, N.C. Nguyen and A.T. Patera *"Natural norm" a posteriori error estimators for reduced basis approximations*. Journal of Computational Physics, (2006).

[47] W. Sickel, *Superposition of functions in Sobolev spaces of fractional order. A survey*. Banach Center Publ. 27, 481-497(1992).

[48] J.C. Slater *A simplification of the Hartree-Fock method*. Phys. Rev., 81(3) :385–390, (1951).

[49] G. Stampacchia, *Le problème de Dirichlet pour les Èquations elliptiques du second ordre à coefficients discontinues*. Ann. Inst. Fourier, tome 15. 189-257 (1965).

[50] G. Strang and G. J. Fix *An analysis of the finite element method.* Prentice-Hall, Englewood Cliffs, N. J., (1973).

[51] L.H. Thomas *The calculation of atomic fields*. Proc. Camb. Phil. Soc., 23 :542–548, (1927).

[52] N. Troullier and J.L. Martins, *A straightforward method for generating soft transferable pseudopotentials*, Solid State Comm. 74. 613-616.(1990).

[53] Y.A. Wang and E.A. Carter, *Orbital-free kinetic energy density functional theory*, in: Theoretical methods in condensed phase chemistry, volume 5 of Progress in theoretical chemistry and physics, pp. 117-184, Kluwer, (2000).

[54] J. Xu *Two-grid discretization techniques for linear and nonlinear PDEs,* SIAM J. Numer. Anal., 33(5):1759-1777, (1996).

[55] J. Xu, *Two-grid discretization techniques for linear and nonlinear PDEs,* SIAM J. Numer. Anal., 33(5):1759-1777, (1996).

[56] J. Xu and A. Zhou, *Local and parallel finite element algorithms based on two-grid discretizations*, Math. Comput., 69, 881-909, (2000).

[57] J. Xu and A. Zhou, *A Two-grid discretization scheme for eigenvalue problems,* Math. Comp., 70: 17-25, (2001).

[58] Xu, J. and A. Zhou. *Local and parallel element algorithms for eigenvalue problems*, Acta Mathematicae Applicates Sinica, English Seris, 18, 185-200 (2002).

[59] A. Zhou *An analysis of finite-dimensional approximations for the ground state solution of Bose–Einstein condensates*, Nonlinearity 17 (2004) 541–550 (2003).

[60] A. Zhou *Finite dimensional approximations for the electronic ground state solution of a molecular system*. Math. Meth. Appl. Sci. x; 30:429–447 (1990)

Résumé

Résumé

Dans ce travail, nous nous intéressons à l'analyse numérique de problèmes aux valeurs propres non linéaires, comme on peut en trouver en chimie quantique ou en mécanique. La résolution de ces problèmes étant très coûteuse, l'idée est de proposer de nouvelles méthodes permettant de simplifier la résolution de ce type de problèmes et ainsi diminuer le coût de calcul. L'analyse numérique est nécessaire pour comprendre si l'impact positif sur le coût de calcul total n'a pas de mauvaise conséquence sur la précision des résultats. On propose un complément aux travaux existants sur les estimations d'erreur *a priori*, afin d'obtenir des résultats équivalents à ceux connus dans le cas de problèmes aux valeurs propres linéaires. Ces résultats ont été utilisés pour la mise en œuvre et l'analyse numérique de nouveaux *schémas à deux grilles* pour l'approximation de problèmes aux valeurs propres non linéaires.
Ensuite, on propose d'adapter ce type de méthode de *sous-grilles*, pour une utilisation associée à la méthode des bases réduites.

Mots-clés : Méthodes à deux grilles, méthodes des bases réduites, éléments finis, ondes planes, problèmes aux valeurs propres, modèle de Thomas-Fermi-Von Weizacker.

Abstract

Abstract

This thesis focuses on the numerical analysis of nonlinear eigenvalue problem, as in quantum chemistry or mechanics. Since solving them is quite expensive. We provice new methods to simplify this computation. The numerical analysis is necessary to understand the effect of this simplification on the accuracy of the result. First, we establish some *a priori* errors estimates of the discretization of for variational approximations of the ground state energy, eigenvalue and eigenvector of nonlinear elliptic eigenvalue problems. Then, we focus on the analysis of the planewave discretization of the Thomas-Fermi-Von-Weizacker model. Since our objectif is to gain in time computation, we provide the numerical analysis of two-grid scheme for a nonlinear eigenvalue problem.
The reduced basis method may be successful in this frame and recent progress have permitted to make the computations reliable thanks to *a posteriori* estimators to extend the method to non linear problems. However, it may not always be possible to use the code to perform all the "off-line" computations required for an efficient performance of the reduced basis method. We propose, in the second part, an alternating approach based on a coarse grid finite element the convergence of which is accelerated through the reduced basis and an improved post processing.

Keywords: Two-grid schemes, reduced basis method, finite elements method, planewaves, nonlinear eigenvalue problem, Thomas-Fermi-Von Weizacker model.