J Sci Comput (2010) 45: 90–117
DOI 10.1007/s10915-010-9358-1

Numerical Analysis of Nonlinear Eigenvalue Problems

Eric Cancès · Rachida Chakir · Yvon Maday

Received: 15 February 2010 / Accepted: 19 February 2010 / Published online: 19 March 2010
© Springer Science+Business Media, LLC 2010

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 Non linear eigenvalue problem · Spectral and pseudo spectral approximations · Finite element approximation · Ground state computations · Numerical analysis

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

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

R. Chakir · Y. Maday (✉)
UMR 7598 LJLL, UPMC Univ Paris 06, Paris 75005, France
e-mail: maday@ann.jussieu.fr

Y. Maday
Division of Applied Mathematics, Brown University, Providence, RI, USA

structure in the framework of nonlinear elasticity, the Gross-Pitaevskii equation describing the steady states of Bose-Einstein condensates [10], 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 [3] for a mathematical introduction).

The numerical analysis of *linear* eigenvalue problems has been thoroughly studied in the past decades (see e.g. [1]). On the other hand, only a few results on *nonlinear* eigenvalue problems have been published so far [14, 15].

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

$$\begin{cases} \Omega \text{ is a regular bounded domain or a rectangular brick of } \mathbb{R}^d \text{ and } X = H_0^1(\Omega) \text{ or} \\ \Omega \text{ is the unit cell of a periodic lattice } \mathcal{R} \text{ of } \mathbb{R}^d \text{ and } X = H_\#^1(\Omega) \end{cases}$$

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

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

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 $k \in \mathbb{N}$,

$$H_\#^s(\Omega) = \left\{ v|_\Omega,\ v \in H_{\rm loc}^s(\mathbb{R}^d) \mid v\ \mathcal{R}\text{-periodic} \right\},$$

$$C_\#^k(\Omega) = \left\{ v|_\Omega,\ v \in C^k(\mathbb{R}^d) \mid v\ \mathcal{R}\text{-periodic} \right\}.$$

We assume in addition that

$$A \in (L^\infty(\Omega))^{d\times d} \text{ and } A(x) \text{ is symmetric for almost all } x \in \Omega; \tag{2}$$

$$\exists \alpha > 0 \text{ s.t. } \xi^T A(x) \xi \geq \alpha |\xi|^2 \text{ for all } \xi \in \mathbb{R}^d \text{ and almost all } x \in \Omega; \tag{3}$$

$$V \in L^p(\Omega) \text{ for some } p > \max(1, d/2); \tag{4}$$

$$F \in C^1([0, +\infty), \mathbb{R}) \cap C^2((0, \infty), \mathbb{R}) \text{ and } F'' > 0 \text{ on } (0, +\infty); \tag{5}$$

$$\exists 0 \leq q < 2,\ \exists C \in \mathbb{R}_+ \text{ s.t. } \forall t \geq 0,\ |F'(t)| \leq C(1 + t^q); \tag{6}$$

$$F''(t)t \text{ locally bounded in } [0, +\infty). \tag{7}$$

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

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

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 repulsive interaction in Bose-Einstein condensates ($m = 2$). These assumptions are also satisfied by the Thomas-Fermi kinetic energy functional (with $m = 5/3$). Although the results contained in this article cannot be straightforwardly applied to electronic structure models (due to the nonlocality of the Coulomb interaction), our arguments can be extended to the case of the Thomas-Fermi-von Weizsäcker model (see [5]).

Remark 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 Gâteaux differentiable 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 (see e.g. [11]). 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

$$\begin{cases} A_u u = \lambda u \\ \|u\|_{L^2} = 1. \end{cases} \tag{11}$$

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

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

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 Sect. 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 Sect. 3, we concentrate on the discretization of problem (1) with

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

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

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

Our results improve those obtained by Zhou in [14], in that we are able to obtain convergence rates for $|\lambda_\delta - \lambda|$, $\|u_\delta - u\|_{H^1}$, $\|u_\delta - u\|_{L^2}$ and $E(u_\delta) - E(u)$. Moreover, the reported numerical test cases tend to show that our error bounds for both Fourier and finite element discretizations, are optimal.

Lastly, we discuss the issue of numerical integration in Sect. 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 at u and $E''(u)$ is the second derivative of E at u.

Lemma 1 *There exist $\beta > 0$ 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

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

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 $v \in X$,

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

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

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

where

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

The existence and uniqueness of the solution to (23) is a straightforward consequence of (18) and the Lax-Milgram lemma. Besides,

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

We can now state the main result of this section.

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

$$\|u_\delta - u\|_{H^1} \underset{\delta \to 0^+}{\longrightarrow} 0. \tag{25}$$

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

$$\frac{\gamma}{2} \|u_\delta - u\|_{H^1}^2 \le E(u_\delta) - E(u) \le \frac{M}{2} \|u_\delta - u\|_{H^1}^2 + C \|u_\delta - u\|_{L^{6/(5-2q)}}, \tag{26}$$

and

$$|\lambda_\delta - \lambda| \le C \left(\|u_\delta - u\|_{H^1}^2 + \|u_\delta - u\|_{L^{6/(5-2q)}} \right). \tag{27}$$

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

$$\|u_\delta - u\|_{H^1} \le C \min_{v_\delta \in X_\delta} \|v_\delta - u\|_{H^1}, \tag{28}$$

$$\|u_\delta - u\|_{L^2}^2 \le C \left(\|u_\delta - u\|_{L^2} \|u_\delta - u\|_{L^{6r/(5-s)}}^r + \|u_\delta - u\|_{H^1} \min_{\psi_\delta \in X_\delta} \|\psi_{u_\delta - u} - \psi_\delta\|_{H^1} \right). \tag{29}$$

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

$$\frac{\gamma}{2} \|u_\delta - u\|_{H^1}^2 \le E(u_\delta) - E(u) \le C \|u_\delta - u\|_{H^1}^2. \tag{30}$$

The result (25) was first established in [14], as well as an inequality similar to (27). On the other hand, the other results seem to be new.

Remark 2 If $0 \le r + s \le 3$, then

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

so that (29) implies the simpler inequality

$$\|u_\delta - u\|_{L^2}^2 \le C \|u_\delta - u\|_{H^1} \min_{\psi_\delta \in X_\delta} \|\psi_{u_\delta - u} - \psi_\delta\|_{H^1}. \tag{31}$$

Proof of Theorem 1 We have

$$E(u_\delta) - E(u) = \frac{1}{2} \langle A_u u_\delta, u_\delta \rangle_{X',X} - \frac{1}{2} \langle A_u u, u \rangle_{X',X}$$

$$+\frac{1}{2}\int_\Omega F\left(u_\delta^2\right)-F\left(u^2\right)-f\left(u^2\right)(u_\delta^2-u^2)$$
$$=\frac{1}{2}\langle (A_u-\lambda)(u_\delta-u),(u_\delta-u)\rangle_{X',X}$$
$$+\frac{1}{2}\int_\Omega F\left(u_\delta^2\right)-F\left(u^2\right)-f\left(u^2\right)(u_\delta^2-u^2). \tag{32}$$

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

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

We then easily deduce from (32) the upper bounds in (26) and (30).

Next, we remark that

$$\lambda_\delta-\lambda=\langle E'(u_\delta),u_\delta\rangle_{X',X}-\langle E'(u),u\rangle_{X',X}$$
$$=a(u_\delta,u_\delta)-a(u,u)+\int_\Omega f(u_\delta^2)u_\delta^2-\int_\Omega f(u^2)u^2$$
$$=a(u_\delta-u,u_\delta-u)+2a(u,u_\delta-u)+\int_\Omega f(u_\delta^2)u_\delta^2-\int_\Omega f(u^2)u^2$$
$$=a(u_\delta-u,u_\delta-u)+2\lambda\int_\Omega u(u_\delta-u)-2\int_\Omega f(u^2)u(u_\delta-u)$$
$$+\int_\Omega f(u_\delta^2)u_\delta^2-\int_\Omega f(u^2)u^2$$
$$=a(u_\delta-u,u_\delta-u)-\lambda\|u_\delta-u\|_{L^2}^2-2\int_\Omega f(u^2)u(u_\delta-u)$$
$$+\int_\Omega f(u_\delta^2)u_\delta^2-\int_\Omega f(u^2)u^2$$
$$=\langle (A_u-\lambda)(u_\delta-u),(u_\delta-u)\rangle_{X',X}+\int_\Omega w_{u,u_\delta}(u_\delta-u) \tag{33}$$

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

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

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

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

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

$$\begin{aligned} |\lambda_\delta - \lambda| &\le M\|u_\delta - u\|_{H^1}^2 + \|w_{u,u_\delta}\|_{L^{6/(2q+1)}} \|u_\delta - u\|_{L^{6/(5-2q)}} \\ &\le M\|u_\delta - u\|_{H^1}^2 + C(1 + \|u_\delta\|_{H^1}^{2q+1}) \|u_\delta - u\|_{L^{6/(5-2q)}} \\ &\le C\left(\|u_\delta - u\|_{H^1}^2 + \|u_\delta - u\|_{L^{6/(5-2q)}}\right), \end{aligned} \tag{34}$$

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

$$\forall v_\delta \in X_\delta, \quad \|u_\delta - u\|_{H^1} \le \|u_\delta - v_\delta\|_{H^1} + \|v_\delta - u\|_{H^1}, \tag{35}$$

and that

$$\begin{aligned} \|u_\delta - v_\delta\|_{H^1}^2 &\le \beta^{-1} \langle (E''(u) - \lambda)(u_\delta - v_\delta), (u_\delta - v_\delta)\rangle_{X',X} \\ &= \beta^{-1}\big(\langle (E''(u) - \lambda)(u_\delta - u), (u_\delta - v_\delta)\rangle_{X',X} \\ &\quad + \langle (E''(u) - \lambda)(u - v_\delta), (u_\delta - v_\delta)\rangle_{X',X}\big). \end{aligned} \tag{36}$$

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

$$\begin{aligned} &\langle (E''(u) - \lambda)(u_\delta - u), w_\delta\rangle_{X',X} \\ &= -\int_\Omega \left(f(u_\delta^2)u_\delta - f(u^2)u_\delta - 2f'(u^2)u^2(u_\delta - u)\right) w_\delta + (\lambda_\delta - \lambda)\int_\Omega u_\delta w_\delta. \end{aligned} \tag{37}$$

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

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

Using (8) and (34), we therefore obtain that for all $0 < \delta \le \delta_1$ and all $v_\delta \in X_\delta$ such that $\|v_\delta\|_{L^2} = 1$,

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

It then follows from (18), (36) and (38) that for all $0<\delta\le\delta_1$ and all $v_\delta\in X_\delta$ such that $\|v_\delta\|_{L^2}=1$,

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

Combining with (35) we obtain that there exist $0<\delta_2\le\delta_1$ and $C\in\mathbb{R}_+$ such that for all $0<\delta\le\delta_2$ and all $v_\delta\in X_\delta$ such that $\|v_\delta\|_{L^2}=1$,

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

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

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

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,

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

For $0<\delta\le\delta_2\le\delta_1$, we have $\|u_\delta^0-u\|_{H^1}\le\|u_\delta-u\|_{H^1}\le 1/2$, and therefore $\|u_\delta^0\|_{H^1}\le\|u\|_{H^1}+1/2$ and $\|u_\delta^0\|_{L^2}\ge 1/2$, yielding $J_\delta\le 2(\|u\|_{H^1}+1)\widetilde{J}_\delta$. Thus (28) is proved.

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

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

from which we infer that

$$\begin{aligned}
\|u_\delta-u\|_{L^2}^2 &= \int_\Omega(u_\delta-u)(u_\delta^*-u)+\int_\Omega(u_\delta-u)(u_\delta-u_\delta^*)\\
&=\int_\Omega(u_\delta-u)(u_\delta^*-u)-\frac{1}{2}\|u_\delta-u\|_{L^2}^2\int_\Omega(u_\delta-u)u\\
&=\int_\Omega(u_\delta-u)(u_\delta^*-u)+\frac{1}{2}\|u_\delta-u\|_{L^2}^2\left(1-\int_\Omega u_\delta u\right)\\
&=\int_\Omega(u_\delta-u)(u_\delta^*-u)+\frac{1}{4}\|u_\delta-u\|_{L^2}^4
\end{aligned}$$

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

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

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

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

$$
\begin{aligned}
&\left| \langle (E''(u) - \lambda)(u_\delta - u), \psi_\delta \rangle_{X',X} \right| \\
&\quad \le C\left(\|u_\delta - u\|_{L^{6r/(5-s)}}^r + \|u_\delta - u\|_{L^{6/5}} \left(\|u_\delta - u\|_{H^1}^2 + \|u_\delta - u\|_{L^{6/(5-2q)}} \right) \right) \|\psi_\delta\|_{H^1}. \qquad (39)
\end{aligned}
$$

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

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

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

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

Lastly, denoting by $\Pi^1_{X_\delta}$ the orthogonal projector on X_δ for the H^1 inner product, a simple calculation leads to

$$\forall v \in u^\perp, \qquad \min_{v_\delta \in X_\delta \cap u^\perp} \|v_\delta - v\|_{H^1} \le \left(1 + \frac{\|\Pi^1_{X_\delta} u\|_{H^1}}{(u, \Pi^1_{X_\delta} u)_{L^2}}\right) \min_{v_\delta \in X_\delta} \|v_\delta - v\|_{H^1}, \tag{40}$$

which completes the proof of Theorem 1. □

Remark 3 In the proof of Theorem 1, we have obtained bounds on $|\lambda_\delta - \lambda|$ from (33), 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

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

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 locally bounded $[0, +\infty)$, then w_{u,u_δ} is uniformly bounded in X. It then follows from (33) that

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

an estimate which is an improvement of (27). 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

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

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

$$\inf\left\{E(v),\ v \in X,\ \int_\Omega v^2 = 1\right\}, \tag{41}$$

where

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

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 (41), 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 [12] that if $\tau > d/2$, $g \in C^{[\tau],\tau-[\tau]+\epsilon}(\mathbb{R},\mathbb{R})$ and $v \in H^\tau_\#(\Omega)$, then $g(v) \in H^\tau_\#(\Omega)$.

A natural discretization of (41) consists in using a Fourier basis. Denoting by $e_k(x) = (2\pi)^{-d/2}e^{ik\cdot x}$, we have for all $v \in L^2(\Omega)$,

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

where $\widehat{v}_k$ is the kth 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 (41) by the Fourier spectral approximation is based on the choice

$$X_\delta = \widetilde{X}_N = \left\{ \sum_{k\in\mathbb{Z}^d\,|\,|k|_*\le N} c_k e_k \mid \forall k,\ c_k^* = c_{-k} \right\},$$

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 $|k|_\infty = \max_{1\le i\le d}|k_i|$). Note that the constraints $c_k^* = c_{-k}$ imply that the functions of $\widetilde{X}_N$ are real-valued. For convenience, the discretization parameter for this approximation will be denoted as N.

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

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

we obtain that for all $\tau \in \mathbb{R}$, and all $v \in H^\tau_\#(\Omega)$, the best approximation of v in $H^\rho_\#(\Omega)$ for any $\rho \le \tau$ is

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

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

$$\forall v \in H^\tau_\#(\Omega), \quad \|v - \Pi_N v\|_{H^\rho} \le \frac{1}{N^{\tau-\rho}}\|v\|_{H^\tau}. \tag{42}$$

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 (42), 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

$$-\Delta u_N + \Pi_N \left[V u_N + f(u_N^2) u_N\right] = \lambda_N u_N. \tag{43}$$

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

$$\begin{aligned} \Delta(u_N - u) &= \Pi_N \left(V(u_N - u) + f(u_N^2)u_N - f(u^2)u\right) \\ &\quad - (I - \Pi_N)(Vu + f(u^2)u) - \lambda_N(u_N - u) - (\lambda_N - \lambda)u. \end{aligned} \tag{44}$$

As $(u_N)_{N\in\mathbb{N}}$ is 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 u 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 (43) that u_N converges to u in $H^{\sigma+2}_\#(\Omega)$.

Besides, the unique solution to (23) solves the elliptic equation

$$\begin{aligned} -\Delta\psi_w &+ \left(V + f(u^2) + 2f'(u^2)u^2 - \lambda\right)\psi_w \\ &= 2\left(\int_\Omega f'(u^2)u^3\psi_w\right) u + w - (w,u)_{L^2} u, \end{aligned} \tag{45}$$

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

$$\|u_N - u\|_{H^\tau} \leq \frac{C}{N^{\sigma+2-\tau}} \quad \text{for } \tau = 0 \text{ and } \tau = 1; \tag{46}$$

$$|\lambda_N - \lambda| \leq \frac{C}{N^{\sigma+2}}; \tag{47}$$

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

From (46) and the inverse inequality

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

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

$$\|u_N - u\|_{H^\tau} \leq \frac{C}{N^{\sigma+2-\tau}} \quad \text{for all } 0 \leq \tau < \sigma + 2. \tag{48}$$

The estimate (47) 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 (33), which we rewrite as,

$$\lambda_N - \lambda = \langle (A_u - \lambda)(u_N - u), (u_N - u)\rangle_{X',X} + \int_\Omega w_{u,u_N}(u_N - u) \tag{49}$$

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 $u/2 \le u_N \le 2u$ 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 (49) that for N large enough

$$|\lambda_N - \lambda| \le C\left(\|u_N - u\|_{H^1}^2 + \|u_N - u\|_{H^{-\sigma}}\right). \tag{50}$$

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

$$\begin{aligned}\int_\Omega w(u_N - u) &= \langle (E''(u) - \lambda)(u_N - u), \Pi^1_{\widetilde{X}_N \cap u^\perp} \psi_w\rangle_{X',X} \\ &\quad + \langle (E''(u) - \lambda)(u_N - u), \psi_w - \Pi^1_{\widetilde{X}_N \cap u^\perp} \psi_w\rangle_{X',X} \\ &\quad + \|u_N - u\|_{L^2}^2 \int_\Omega f'(u^2) u^3 \psi_w - \frac{1}{2}\|u_N - u\|_{L^2}^2 \int_\Omega uw,\end{aligned} \tag{51}$$

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 (45) that ψ_w is in $H_\#^{\rho+2}(\Omega)$ and satisfies

$$\|\psi_w\|_{H^{\rho+2}} \le C\|w\|_{H^\rho}, \tag{52}$$

for some constant C independent of w.

Combining (18), (39), (40), (48), (49), (51) and (52), we obtain that there exists a constant $C \in \mathbb{R}_+$ such that for all $N \in \mathbb{N}$ and all $w \in H_\#^r(\Omega)$,

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

Therefore

$$\|u_N - u\|_{H^{-\rho}} = \sup_{w \in H_\#^\rho(\Omega)\setminus\{0\}} \frac{\int_\Omega w(u_N - u)}{\|w\|_{H^\rho}} \le \frac{C}{N^{\sigma+2+\rho}}, \tag{53}$$

for some constant $C \in \mathbb{R}_+$ independent of N. Using (48) and (50), 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 $N \in \mathbb{N}$,*

$$\|u_N - u\|_{H^\tau} \leq \frac{C}{N^{\sigma+2-\tau}} \quad \textit{for all } -\sigma \leq \tau < \sigma + 2; \tag{54}$$

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

$$\frac{\gamma}{2}\|u_N - u\|^2_{H^1} \leq E(u_N) - E(u) \leq C\|u_N - u\|^2_{H^1}. \tag{55}$$

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

$$|\lambda_N - \lambda| \leq \frac{C}{N^{2(\sigma+1)}}. \tag{56}$$

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

$$\widehat{V}_k = -\frac{1}{\sqrt{2\pi}}\frac{1}{|k|^2 - \frac{1}{4}}, \tag{57}$$

from which we deduce that $V \in H^\sigma_\#(0, 2\pi)$ for all $\sigma < 3/2$. It can be see on Fig. 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 (54) (for $s = 1$ and $s = 0$), (53) (for $r = 1$) and (56), 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.

4 Finite Element Discretization

In this section, we consider the problem

$$\inf\left\{E(v),\ v \in X,\ \int_\Omega v^2 = 1\right\}, \tag{58}$$

where

$$\Omega \text{ is a rectangular brick of } \mathbb{R}^d, \quad \text{with } d = 1, 2 \text{ or } 3,$$

$$X = H^1_0(\Omega),$$

$$E(v) = \frac{1}{2}\int_\Omega |\nabla v|^2 + \frac{1}{2}\int_\Omega V v^2 + \frac{1}{2}\int_\Omega F(v^2).$$

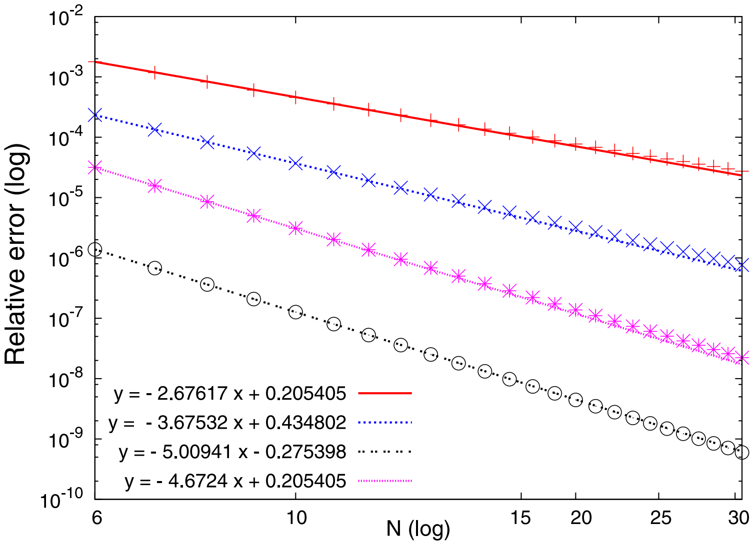

Fig. 1 Numerical errors $\|u_N - u\|_{H^1}$ (+), $\|u_N - u\|_{L^2}$ (×), $\|u_N - u\|_{H^{-1}}$ (∗), and $|\lambda_N - \lambda|$ (○), as functions of $2N + 1$ (the dimension of $\widetilde{X}_N$) in log scales

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 (58) 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 quasi-uniform 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

1. $\overline{\Omega}$ is the union of all the elements of $\mathcal{T}_h$;
2. 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;
3. 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 and K;
4. there exists a constant $C \geq 1$ independent of h and K such that $h_K \leq h \leq Ch_K$ for any h and $K \in \mathcal{T}_h$.

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 $X_{h,k} \subset H_0^1(\Omega)$ as soon as $k \geq 1$.

We denote by $\pi_{h,k}^0$ and $\pi_{h,k}^1$ 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. [8]): there exists $C \in \mathbb{R}_+$ such that for all $r \in \mathbb{N}$ such that $1 \leq r \leq k + 1$,

$$\forall \phi \in H^r(\Omega) \cap H_0^1(\Omega), \quad \|\phi - \pi_{h,k}^0 \phi\|_{L^2} \leq Ch^r \|\phi\|_{H^r},$$

$$\forall \phi \in H^r(\Omega) \cap H_0^1(\Omega), \quad \|\phi - \pi_{h,k}^1 \phi\|_{H^1} \le C h^{r-1} \|\phi\|_{H^r}. \tag{59}$$

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

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

such that $(u_{h,k}, u)_{L^2} \ge 0$. 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 $1 < r \le 2$ and $0 \le r + s \le 3$. Then there exist $h_0 > 0$ and $C \in \mathbb{R}_+$ such that for all $0 < h \le h_0$,*

$$\|u_{h,1} - u\|_{H^1} \le Ch \tag{60}$$

$$\|u_{h,1} - u\|_{L^2} \le Ch^2 \tag{61}$$

$$|\lambda_{h,1} - \lambda| \le Ch^2 \tag{62}$$

$$\frac{\gamma}{2} \|u_{h,1} - u\|_{H^1}^2 \le E(u_{h,1}) - E(u) \le Ch^2. \tag{63}$$

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 locally bounded in $[0, +\infty)$, then there exist $h_0 > 0$ and $C \in \mathbb{R}_+$ such that for all $0 < h \le h_0$,*

$$\|u_{h,2} - u\|_{H^1} \le Ch^2 \tag{64}$$

$$\|u_{h,2} - u\|_{L^2} \le Ch^3 \tag{65}$$

$$|\lambda_{h,2} - \lambda| \le Ch^4 \tag{66}$$

$$\frac{\gamma}{2} \|u_{h,2} - u\|_{H^1}^2 \le E(u_{h,2}) - E(u) \le Ch^4. \tag{67}$$

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

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

to establish that $\psi_{u_{h,k}-u} \in H^2(\Omega) \cap H_0^1(\Omega)$ and that

$$\|\psi_{u_{h,k}-u}\|_{H^2} \le C \|u_{h,k} - u\|_{L^2} \tag{68}$$

for some constant C independent of h and k. The estimates (60)–(63) then are directly consequences of Theorem 1, (31), (59) and (68).

Under the additional assumptions that $V \in H^1(\Omega)$, we obtain by standard elliptic regularity arguments that $u \in H^3(\Omega)$. This follows from a prolongation by reflection argument using the fact that u, hence $\lambda u - Vu - f(u^2)u$, vanish on the boundary $\partial\Omega$ of the rectangular brick Ω. As $u \in H_0^1(\Omega)$ and $-\Delta u = \lambda u - Vu - f(u^2)u \in H_0^1(\Omega)$, the prolongation by reflection $\widetilde{u}$ of u is in $H_\#^1(2\Omega)$ and is such that $-\Delta \widetilde{u} \in H_\#^1(2\Omega)$. Hence $\widetilde{u} \in H_\#^3(2\Omega)$, and therefore $u \in H^3(\Omega)$.

The H^1 and L^2 estimates (64) and (65) immediately follows from Theorem 1, (31), (59) and (68). We also have

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

for a constant C independent of h. In order to prove (66), we proceed as in Sect. 3. 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

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

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

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

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

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

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 [8] for instance), we obtain

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

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 $0 < h \le h_0$,

$$|\lambda_{h,2} - \lambda| \le C\left(\|u_{h,2} - u\|_{H^1}^2 + \|u_{h,2} - u\|_{H^{-1}}\right). \tag{69}$$

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

$$
\begin{aligned}
&-\Delta\psi_w + (V + f(u^2) + 2f'(u^2)u^2 - \lambda)\psi_w \\
&\quad = 2\left(\int_\Omega f'(u^2)u^3\psi_w\right)u + w - (w,u)_{L^2}u,
\end{aligned}
\tag{70}
$$

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 locally bounded in $[0,+\infty)$, we deduce from (70) that ψ_w is in $H^3(\Omega)$ and that there exists $C \in \mathbb{R}_+$ such that for all $w \in H^1_0(\Omega)$ and all $0 < h \le h_0$,

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

We therefore obtain the inequality

$$\|\psi_w - \pi^1_{h,2}\psi_w\|_{H^1} \le Ch^2\|w\|_{H^1}, \tag{71}$$

where the constant C is independent of h.

Putting together (8) (for $r = 2$), (18), (39), (40), (59), (64), (65) and (71), we get

$$\|u_{h,2} - u\|_{H^{-1}} = \sup_{w\in H^1_0(\Omega)\setminus\{0\}} \frac{\int_\Omega w(u_{h,2}-u)}{\|w\|_{H^1}} \le Ch^4.$$

Combining with (64) and (69), we end up with (66). Lastly, we deduce (67) 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 local boundedness of the function $F''(t)t^{1/2}$ in $[0,+\infty)$. □

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 Fig. 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.

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 is the identity matrix.

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 [2, 4, 7]. At the pth iteration ($p \ge 1$), 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

$$\forall v_\delta \in X_\delta, \quad \int_\Omega \overline{\nabla u_\delta^p}\cdot\nabla v_\delta + \int_\Omega \left(V + f(\widetilde{\rho}_\delta^{p-1})\right)\overline{u_\delta^p}\, v_\delta = \lambda_\delta^p \int_\Omega \overline{u_\delta^p} v_\delta, \tag{72}$$

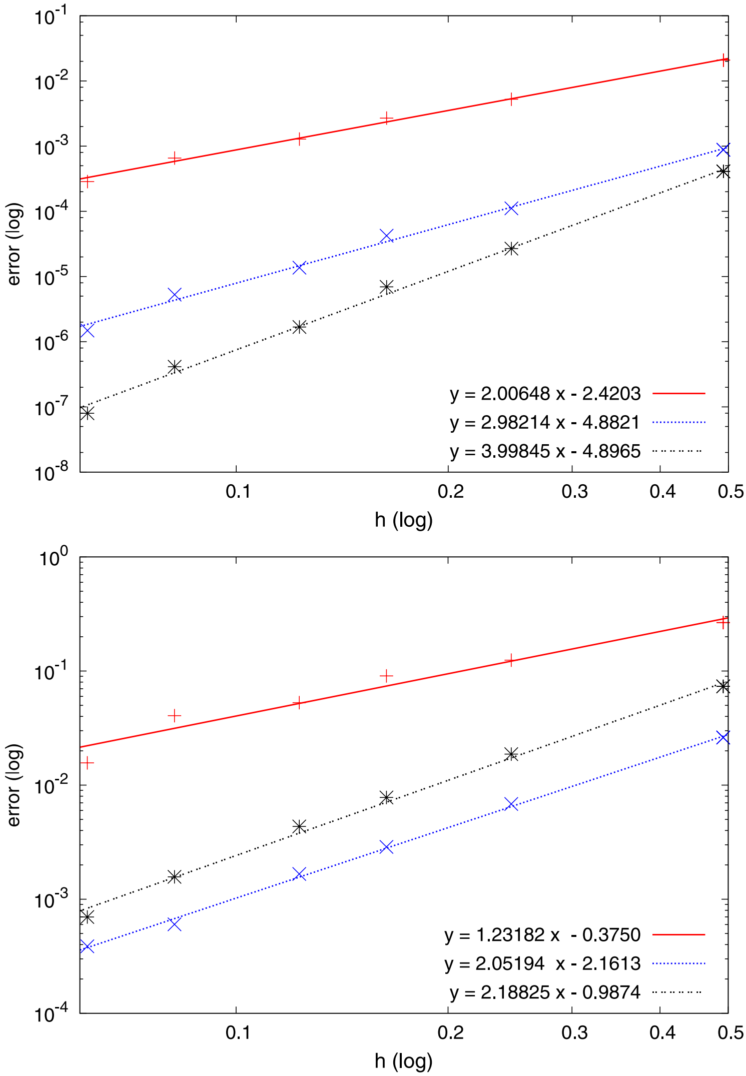

Fig. 2 Errors $\|u_{h,k}-u\|_{H^1}$ (+), $\|u_{h,k}-u\|_{L^2}$ (×) and $|\lambda_{h,k}-\lambda|$ (∗) for the $\mathbb{P}_1$ ($k=1$, *top*) and $\mathbb{P}_2$ ($k=2$, *bottom*) approximations as a function of h in log scales

has to be computed. In the optimal damping algorithm, the density $\widetilde{\rho}_\delta^{p-1}$ is a convex linear combination of the densities $\rho_\delta^q=|u_\delta^q|^2$, for $0\le q\le p-1$. Solving (72) amounts to finding the lowest eigenelement of the matrix H^p with entries

$$H_{kl}^p := \int_\Omega \overline{\nabla\phi_k}\cdot\nabla\phi_l + \int_\Omega V\,\overline{\phi_k}\,\phi_l + \int_\Omega f(\widetilde{\rho}_\delta^{p-1})\,\overline{\phi_k}\,\phi_l, \tag{73}$$

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

In order to evaluate the last two terms of the right-hand side of (73), numerical integration has to be resorted to. In the finite element approximation of (58), 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 (41), 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 $(k-l)$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

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

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 4 In the case of the periodic problem considered in Sect. 3 and when $F(t)=ct^2$ for some $c>0$, the last term of the right-hand side of (73) 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 $\mathrm{Span}\{e_n\,|\,|n|_*\le 4N\}$. 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 [5].

In view of Remark 4, we consider an integration grid

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

with $N_g\ge 4N+1$ 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 $\rho\in\widetilde{X}_{2N}$,

$$\forall|k|,|l|\le N,\quad \int_0^{2\pi}\rho\,\overline{e_k}\,e_l=\frac{1}{N_g}\sum_{r\in\frac{2\pi}{N_g}\mathbb{Z}\cap[0,2\pi)}\rho(r)e^{-i(k-l)r}=\widehat{\rho}^{\mathrm{FFT}}_{k-l}, \tag{74}$$

where $\widehat{\rho}^{\mathrm{FFT}}_{k-l}$ is the $(k-l)$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 $(\widehat{\phi_g^{\rm FFT}})_{g\in\mathbb{Z}}$ defined by

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

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 $W_M=\widetilde{X}_{M/2-1}\oplus\mathbb{C}(e_{M/2}+e_{-M/2})$ if M is even (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

$$\mathcal{I}_{N_g}(\phi)=\begin{cases}(2\pi)^{1/2}\sum_{|g|\le(N_g-1)/2}\widehat{\phi_g^{\rm FFT}}\,e_g & (N_g\text{ odd}),\\ (2\pi)^{1/2}\sum_{|g|\le N_g/2-1}\widehat{\phi_g^{\rm FFT}}\,e_g+(2\pi)^{1/2}\widehat{\phi_{N_g/2}^{\rm FFT}}(\frac{e_{N_g/2}+e_{-N_g/2}}{2}) & (N_g\text{ even}).\end{cases}$$

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

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

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

$$\begin{aligned}H^p_{kl}&:=\int_0^{2\pi}\overline{e_k'}\cdot e_l'+\int_0^{2\pi}V\,\overline{e_k}\,e_l+\int_0^{2\pi}\widetilde{\rho}_N^{p-1}\overline{e_k}\,e_l\\ &\simeq |k|^2\delta_{kl}+\widehat{V_{k-l}^{\rm FFT}}+\widehat{[\widetilde{\rho}_N^{p-1}]_{k-l}^{\rm FFT}},\end{aligned}\tag{75}$$

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 (74)). Of course, this situation is specific to the nonlinearity $F(t)=t^2/2$ considered here.

Using the approximation formula (75) amounts to replace the original problem

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

with the approximate problem

$$\inf\left\{E_{N_g}(v_N),\ v_N\in\widetilde{X}_N,\ \int_0^{2\pi}|v_N|^2=1\right\},\tag{77}$$

where

$$E_{N_g}(v_N)=\frac12\int_0^{2\pi}|v_N'|^2+\frac12\int_0^{2\pi}\mathcal{I}_{N_g}(V)v_N^2+\frac14\int_0^{2\pi}|v_N|^4.$$

Let us denote by u_N a solution of (76) such that $(u_N, u)_{L^2} \geq 0$ and by u_{N,N_g} a solution to (77) 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 $L^\infty_\#(0, 2\pi)$, 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 $(|u_{N,N_g}|^2 - |u_N|^2) \in \widetilde{X}_{2N}$. Therefore,

$$\|u_{N,N_g} - u_N\|_{H^1} \leq C\|\Pi_{2N}(V - \mathcal{I}_{N_g}(V))\|_{L^2}, \tag{78}$$

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

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

from which we deduce, using (78),

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

An error analysis of the interpolation operator $\mathcal{I}_{N_g}$ is given in [6]: for all non-negative real numbers $0 \leq r \leq s$ with $s > 1/2$ (for $d = 1$),

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

Thus,

$$\|\Pi_{2N}(V - \mathcal{I}_{N_g}(V))\|_{L^2} \leq \|V - \mathcal{I}_{N_g}(V)\|_{L^2} \leq \frac{C}{N_g^\sigma}, \tag{79}$$

and the above inequality provides the following estimates:

$$\|u_{N,N_g} - u\|_{H^1} \leq C\left(N^{-\sigma-1} + N_g^{-\sigma}\right), \tag{80}$$

$$\|u_{N,N_g} - u\|_{L^2} \leq C\left(N^{-\sigma-2} + N_g^{-\sigma}\right), \tag{81}$$

$$|\lambda_{N,N_g} - \lambda| \leq C\left(N^{-2\sigma-2} + N_g^{-\sigma}\right), \tag{82}$$

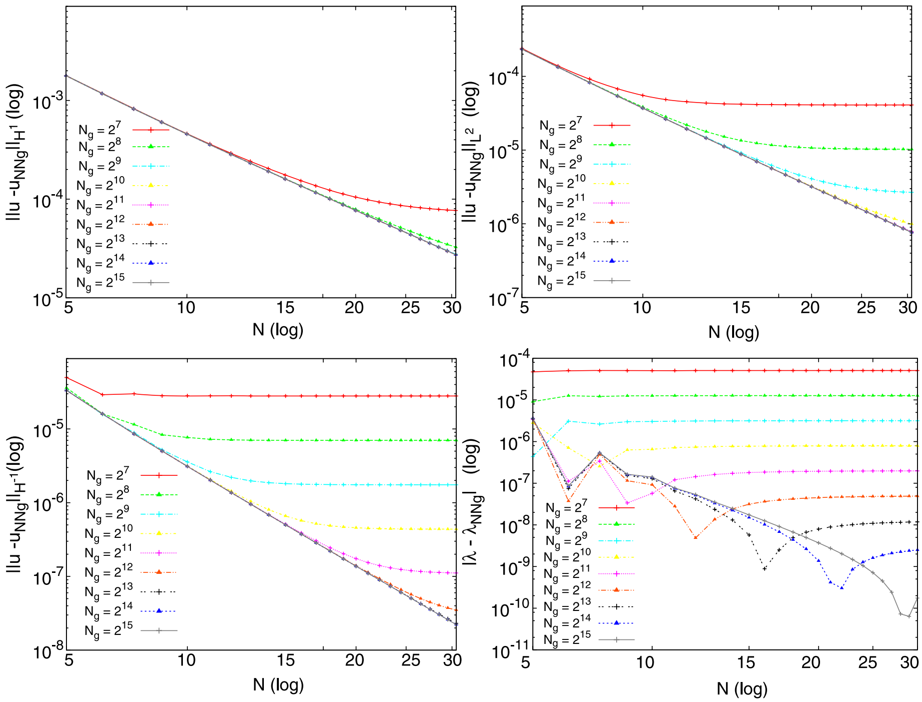

Fig. 3 (Color online) Numerical errors $\|u_{N,N_g} - u\|_{H^1}$ (*top left*), $\|u_{N,N_g} - u\|_{L^2}$ (*top right*), $\|u_{N,N_g} - u\|_{H^{-1}}$ (*bottom left*), and $|\lambda_{N,N_g} - \lambda|$ (*bottom right*), as functions of $2N+1$ (the dimension of $\widetilde{X}_N$), for $N_g = 2^7 = 128$ (*red*), $N_g = 2^8 = 256$ (*green*), $N_g = 2^9 = 512$ (*cyan*), $N_g = 2^{10} = 1{,}024$ (*gold*), $N_g = 2^{11} = 2{,}048$ (*magenta*), $N_g = 2^{12} = 4{,}096$ (*orange*), $N_g = 2^{13} = 8{,}192$ (*black*), $N_g = 2^{14} = 16{,}384$ (*blue*), $N_g = 2^{15} = 32{,}768$ (*light grey*)

for a constant C independent of N and N_g. The first component of the error bound (80) 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 (81) and (82)).

It is classical that 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 (47) 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 Sect. 3 ($\Omega = (0, 2\pi)$, $V(x) = \sin(|x-\pi|/2)$, $F(t) = t^2/2$), we have computed for $4 \le N \le 30$ and $N_g = 2^p$ with $7 \le p \le 15$, 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 Fig. 3, these quantities are plotted as functions of $2N+1$ (the dimension of $\widetilde{X}_N$), for various values of N_g.

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 Fig. 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

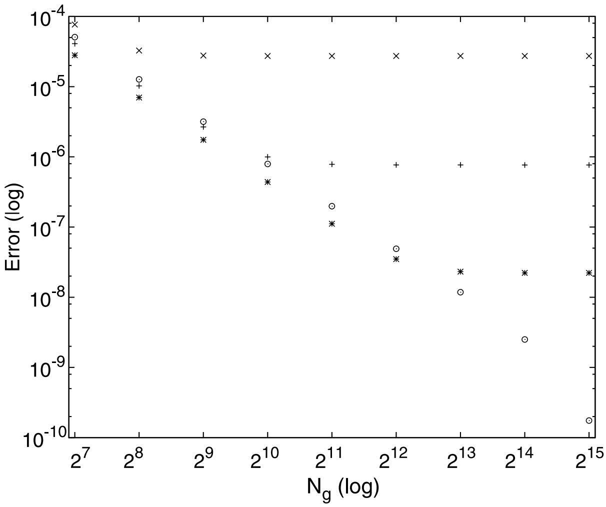

Fig. 4 Numerical errors $\|u_{N,N_g} - u\|_{H^1}$ ($\times$), $\|u_{N,N_g} - u\|_{L^2}$ ($+$), $\|u_{N,N_g} - u\|_{H^{-1}}$ ($*$), and $|\lambda_{N,N_g} - \lambda|$ ($\circ$), for $N = 30$, as functions of N_g (in log scales)

$\log_{10}|\lambda_{N,N_g} - \lambda|$ 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 (80)–(82) 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 (79) 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 (57), show that

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

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

$$\|u_{N,N_g} - u\|_{H^1} \leq C\big(N^{-5/2} + N^{1/2}N_g^{-2}\big),$$
$$\|u_{N,N_g} - u\|_{L^2} \leq C\big(N^{-7/2} + N^{1/2}N_g^{-2}\big),$$
$$|\lambda_{N,N_g} - \lambda| \leq C\big(N^{-5} + N^{1/2}N_g^{-2}\big).$$

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.

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 $V \in L^p(\Omega)$ for some $p > \max(1, d/2)$, $v \mapsto a(v, v)$ 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

$$\forall v \in X, \quad a(v, v) \geq \|v\|_{L^2}^2 \quad \text{and} \quad F(0) = F'(0) = 0. \tag{83}$$

It then follows from (6) and (83) 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 Gâteaux differentiable at u 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 [9], 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 [13]. 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 $\{\rho \geq 0 \mid \sqrt{\rho} \in X, \int_\Omega \rho = 1\}$. 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

$$\lambda_0 = \inf\left\{\int_\Omega (A\nabla v)\cdot\nabla v + \int_\Omega (V + f(u^2))v^2,\ v \in X,\ \int_\Omega v^2 = 1\right\}. \tag{84}$$

Let v_0 be a normalized eigenvector of A_u associated with λ_0. Clearly, v_0 is a minimizer of (84) 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

$$\begin{cases} \text{search } (\mu, v) \in \mathbb{R} \times X \text{ such that} \\ A_v v = \mu v \\ \|v\|_{L^2} = 1, \end{cases} \tag{85}$$

in the following sense: if (μ, v) is solution to (85) 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 (85) and denote by $\widetilde{w} = |v| - u$. As for u, we infer from elliptic regularity arguments [9] that $v \in C^{0,\alpha}(\overline{\Omega})$. We have $\|v\|_{L^2} = \|u\|_{L^2} = 1$. Therefore, if $w \le 0$ 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

$$(A_u - \lambda) w + \frac{f(v^2) - f(u^2)}{v^2 - u^2} v(u + v) w = (\mu - \lambda) v. \tag{86}$$

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 (86) that

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

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

References

1. Babuška, I., Osborn, J.: Eigenvalue problems. In: Handbook of Numerical Analysis. vol. II, pp. 641–787. North-Holland, Amsterdam (1991)
2. Cancès, E.: SCF algorithms for Kohn-Sham models with fractional occupation numbers. J. Chem. Phys. **114**, 10616–10623 (2001)
3. Cancès, E., Defranceschi, M., Kutzelnigg, W., Le Bris, C., Maday, Y.: Computational quantum chemistry: a primer. In: Ciarlet, Ph., Le Bris, C. (eds.) Handbook of Numerical Analysis, pp. 3–270. North-Holland, Amsterdam (2003). Special volume: Computational Chemistry, vol. X
4. Cancès, E., Le Bris, C.: Can we outperform the DIIS approach for electronic structure calculations? Int. J. Quantum Chem. **79**, 82–90 (2000)
5. Cancès, E., Chakir, R., Maday, Y.: Numerical analysis of the planewave discretization of some orbital-free and Kohn-Sham models. In preparation
6. Canuto, C., Hussaini, M.Y., Quarteroni, A., Zang, T.A.: Spectral Methods. Springer, Berlin (2007)
7. Dion, C., Cancès, E.: Spectral method for the time-dependent Gross-Pitaevskii equation with harmonic traps. Phys. Rev. E **67**, 046706 (2003)
8. Ern, A., Guermond, J.-L.: Theory and Practice of Finite Elements. Springer, Berlin (2004)
9. Gilbarg, D., Trudinger, N.S.: Elliptic Partial Differential Equations of Second Order, 3rd edn. Springer, Berlin (1998)
10. Pitaevskii, L.P., Stringari, S.: Bose-Einstein Condensation. Clarendon, Oxford (2003)
11. Reed, M., Simon, B.: Methods of Modern Mathematical Physics, vol. I. Academic Press, New York (1975)
12. Sickel, W.: Superposition of functions in Sobolev spaces of fractional order. A survey. Banach Center Publ. **27**, 481–497 (1992)
13. Stampacchia, G.: Le probléme de Dirichlet pour les équations elliptiques du second ordre á coefficients discontinues. Ann. Inst. Fourier **15**, 189–257 (1965)
14. Zhou, A.: An analysis of finite-dimensional approximations for the ground state solution of Bose-Einstein condensates. Nonlinearity **17**, 541–550 (2004)
15. Zhou, A.: Finite dimensional approximations for the electronic ground state solution of a molecular system. Math. Meth. Appl. Sci. **30**, 429–447 (2007)