GLOBAL CONVERGENCE OF AN EFFICIENT SPLITTING METHOD FOR THE DEFOCUSING GROSS–PITAEVSKII GROUND STATE PROBLEM

JIAXING LI, SHIXIN ZHENG, AND XIANGXIONG ZHANG

Abstract. For computing the ground state of the defocusing Gross–Pitaevskii energy, we propose and analyze two efficient schemes based on the Davis–Yin three-operator splitting, which treats the potential and interaction terms explicitly, and the kinetic energy by a resolvent. One iteration costs one inversion of $\mathbb{I} - \gamma\Delta$ for the first scheme, and of $\mathbb{I} - \gamma\Delta + \gamma V_1$ for the second, with V_1 denoting the separable part of the potential, and on structured meshes both operators can be inverted by simple fast GPU solvers. For monotone discrete Laplacians, including the second-order finite difference scheme and the lumped linear finite element method on simplicial meshes with suitable angle conditions, we prove global convergence to the unique positive discrete ground state, for every positive normalized initial vector, for any constant step size below an explicit threshold. In contrast, the methods previously proven to converge globally to the ground state all invert a more difficult elliptic operator that depends on the iteration variable. In three-dimensional tests with up to 999^3 unknowns on one GPU, a simple variable step size rule makes the proposed splitting schemes efficient in practice, comparable in wall-clock time to Riemannian conjugate gradient methods that also invert only a shifted Laplacian operator, and much more robust with respect to the choice of the initial guess.

Keywords: Gross–Pitaevskii equation, ground state, Davis–Yin splitting, global convergence

2020 Mathematics Subject Classification: Primary 65N25, 65N12, 65N30, 49R05; Secondary 65K99, 49Q99, 90-08, 90C26.

1. Introduction

1.1. The Gross–Pitaevskii ground state problem.

A standard model for the equilibrium states of a Bose–Einstein condensate [31] is the minimization of the Gross–Pitaevskii (GP) energy. With a standard rescaling (see, e.g., [28, 5]) the GP ground state problem on a bounded domain $\Omega \subset \mathbb{R}^d$, for instance $\Omega = [-L, L]^d$, can be stated as minimizing the energy functional

$$E(\phi) = \frac{1}{2}\int_\Omega \left(|\nabla\phi(\boldsymbol{x})|^2 + V(\boldsymbol{x})|\phi(\boldsymbol{x})|^2\right) \mathrm{d}\boldsymbol{x} + \frac{\beta}{4}\int_\Omega |\phi(\boldsymbol{x})|^4 \,\mathrm{d}\boldsymbol{x}, \tag{1.1}$$

over the constraint set

$$\left\{\phi \in H_0^1(\Omega) : \int_\Omega |\phi(\boldsymbol{x})|^2 \,\mathrm{d}\boldsymbol{x} = 1\right\}, \tag{1.2}$$

where $d = 1, 2, 3$ is the dimension, $V(\boldsymbol{x})$ is a given potential, and β is the coupling constant. We only consider the case $\beta > 0$, which is the *defocusing* (repulsive) regime.

Date: September 11, 2026.

J. Li and X. Zhang: Department of Mathematics, Purdue University, 150 N. University Street, West Lafayette, IN 47907 (`li4944@purdue.edu`, `zhan1966@purdue.edu`).

S. Zheng: Department of Mathematics, University of Maryland, College Park, MD 20742 (`siuzheng@umd.edu`).

XZ was supported in part by NSF DMS-2208518.

The existence, uniqueness, and regularity of the GP ground state for $\beta > 0$ are well understood, see, e.g., [28]. For $\beta > 0$ the minimizer of (1.1)–(1.2) is unique up to sign, and its positive representative $u^* > 0$, called *the* ground state, is an eigenfunction of the nonlinear eigenvalue problem

$$-\Delta u(\boldsymbol{x}) + V(\boldsymbol{x})u(\boldsymbol{x}) + \beta|u(\boldsymbol{x})|^2 u(\boldsymbol{x}) = \lambda u(\boldsymbol{x}), \qquad \int_\Omega |u(\boldsymbol{x})|^2 \, \mathrm{d}\boldsymbol{x} = 1, \qquad u(\boldsymbol{x})|_{\partial\Omega} = 0. \tag{1.3}$$

Since a constant shift of the potential changes the energy (1.1) only by a constant over the constraint set (1.2), the assumption $V \geq 0$ loses no generality. Here (1.3) is understood in the sense of distributions, i.e., in the variational form: seek $\lambda \in \mathbb{R}$ and $u \in H_0^1(\Omega)$ satisfying

$$(\nabla u, \nabla v) + (Vu, v) + \beta(|u|^2 u, v) = \lambda\,(u, v), \qquad \forall\, v \in H_0^1(\Omega), \tag{1.4}$$

where $(u, v) = \int_\Omega u(\boldsymbol{x})v(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x}$. Taking $u = v = u^*$ in (1.4) shows that the ground state eigenvalue is determined by the ground state through

$$\lambda^* = 2E(u^*) + \frac{\beta}{2}\int_\Omega |u^*(\boldsymbol{x})|^4 \, \mathrm{d}\boldsymbol{x}. \tag{1.5}$$

1.2. **Numerical methods and their convergence theory.** After a spatial discretization, the ground state can be solved by minimizing the discrete GP energy over the convex sphere constraint in $\mathbb{R}^N$. We refer to the survey [5] for models and methods, and to the recent review [23] for a systematic account of the convergence theory, whose terminology we follow. *Algebraic methods* discretize first and then iterate on the finite-dimensional problem: the self-consistent field (SCF) iteration [8, 10, 34], the inverse iteration or $\mathscr{A}$-method [25], and the J-method [1], which linearizes with the Jacobian of the nonlinear eigenvalue problem and admits spectral shifts. *Variational methods* discretize an iteration derived in function space: the gradient flow with discrete normalization (GFDN) [6], a semi-implicit backward Euler step for the projected L^2-gradient flow and the Sobolev gradient flows obtained by representing the gradient in other metrics, namely the H^1- and a_0-flows [26, 15, 16, 11] and the energy-adaptive a_u-flow [24, 2]. Riemannian optimization with conjugate gradient acceleration [16, 4] and Newton-type methods [36, 3] are faster in practice, but the conjugate gradient variants come without a convergence theory [23, §5.3], apart from the discussion in [4, §5.4], which sketches an argument for convergence to a critical point and gives heuristics for the rate, and among the Newton-type methods the available guarantee is the convergence of the residual to zero for the regularized Newton method [36], by a trust-region argument, which yields a stationary point rather than the ground state.

Convergence of these methods can be summarized into three levels.

(i) Local convergence to the ground state. Many methods have been proven to converge locally at a linear rate governed by a spectral gap of a linearized operator, such as SCF [34], the $\mathscr{A}$-method [25], the J-method [1], and the Sobolev gradient flows [39, 24, 22, 11, 12].

(ii) Global convergence to a critical point. From any initial guess, a monotone energy decay combined with a summability argument shows that every accumulation point is a constrained critical point of the energy, as shown for the H^1- and a_0-flows [11, 12]. For the rotating energy functional and for two-component condensates, energy dissipation and global convergence to a critical point were established in [19, 38, 18].

(iii) Global convergence to the ground state. The few results at this level all rest on positivity as a selection mechanism. Henning and Peterseim [24] proved that the a_u-flow started from a nonnegative initial guess converges in H^1 to a strictly positive eigenfunction, necessarily the ground state, since the Gross–Pitaevskii energy admits no positive excited state. The same conclusion was

reached for the damped J-method [1, Prop. 3]. For the classical GFDN even the energy-diminishing property was a long-standing open problem [23, §5.2.1]. It was settled recently by Feng, Tang and Wang [20], together with global convergence to a critical point, and in the defocusing case without rotation the whole sequence started from a nonnegative initial value converges in H^1 to the ground state, by the positivity preservation of the semi-implicit step [20, Cor. 1]. For the continuous-in-time L^2-normalized gradient flow, which is the model behind the GFDN rather than the algorithm itself, convergence to the ground state from a nonnegative initial value was proven recently in [13]. These arguments use the uniqueness and positivity of the *continuous* ground state, properties that a spatial discretization does not automatically inherit. They are, however, available for the monotone discretizations of Section 2. For the lumped P^1 and finite difference schemes, the discrete ground state is unique up to sign, strictly positive, and the only nonnegative critical point of the discrete energy on the sphere, by the M-matrix structure combined with either the Perron–Frobenius theorem [12] or a discrete Picone inequality [21]. With these properties, the fully discrete a_u-flow for the lumped P^1 scheme was proven to converge to the discrete ground state for every nonnegative normalized initial guess [21, Cor. 4.2].

1.3. **The scheme and its computational cost.** The three methods that reach level (iii), the a_u-flow, the damped J-method and the GFDN, all invert an elliptic operator that depends on the iteration variable, and no method whose iteration inverts only a shifted Laplacian has been proven to converge globally to the ground state, e.g., it is quite difficult to prove that the H^1- or a_0-scheme can preserve positivity [11, §3.1]. Motivated by this gap, we propose and analyze a semi-implicit splitting method for the discrete ground state problem (2.5) and prove its global convergence to the unique positive discrete ground state. By applying the Davis–Yin three-operator splitting [17] to natural three-part decompositions of (2.5), and by derivation in Appendix B, with notation in Section 2, we obtain two schemes:

$$
\begin{cases}
\mathbf{u}^k = \dfrac{\mathbf{z}^k}{\|\mathbf{z}^k\|_2}, \\
\mathbf{x}^k = (\mathbb{I} - \gamma\Delta_h)^{-1}\big(2\mathbf{u}^k - \mathbf{z}^k - \gamma(\mathbb{V}\mathbf{u}^k + \beta(\mathbf{u}^k)^3)\big), \\
\mathbf{z}^{k+1} = \mathbf{z}^k + \mathbf{x}^k - \mathbf{u}^k,
\end{cases}
\tag{1.6}
$$
(DYS I)

$$
\begin{cases}
\mathbf{u}^k = \dfrac{\mathbf{z}^k}{\|\mathbf{z}^k\|_2}, \\
\mathbf{x}^k = (\mathbb{I} - \gamma\Delta_h + \gamma\mathbb{V}_1)^{-1}\big(2\mathbf{u}^k - \mathbf{z}^k - \gamma(\mathbb{V}_2\mathbf{u}^k + \beta(\mathbf{u}^k)^3)\big), \\
\mathbf{z}^{k+1} = \mathbf{z}^k + \mathbf{x}^k - \mathbf{u}^k,
\end{cases}
\tag{1.7}
$$
(DYS II)

where $\gamma > 0$ is the step size, $V = V_1 + V_2$ with $V_1 \geq 0$ and $V_2 \geq 0$, $\mathbb{V}_1$ and $\mathbb{V}_2$ are the corresponding diagonal matrices, and Δ_h denotes the discrete Laplacian.

We refer to the first scheme (1.6) as DYS I in which only a shifted Laplacian is inverted. When V_1 is *separable*, i.e., a sum of one-dimensional potentials, each depending on a single coordinate x_j, as is the trapping potential $|\boldsymbol{x}|^2 = \sum_{j=1}^d x_j^2$, we can also consider the second scheme (1.7), referred to as DYS II, in which $(\mathbb{I} - \gamma\Delta_h + \gamma\mathbb{V}_1)^{-1}$ can still be inverted easily and efficiently on GPUs on structured meshes, by the same per-axis eigendecomposition that inverts $-\Delta_h$ directly, see [30].

From an implementation point of view, the methods reviewed in §1.2 differ mainly in the linear system solved at each iteration, compared in Table 1. In general, a shifted Laplacian is much easier to invert than an elliptic operator containing a spatially varying potential term.

method	operator to invert	solves/iter.	convergence proved
GFDN	$-\Delta_h + \mathbb{V} + \beta\,\mathrm{diag}(\mathbf{u}^2) + \tau^{-1}\mathbb{I}$	1	ground state [20]
damped J	$-\Delta_h + \mathbb{V} + 3\beta\,\mathrm{diag}(\mathbf{u}^2) - \sigma\mathbb{I}$ (+ rank one)	2	ground state [1]
a_u-flow	$-\Delta_h + \mathbb{V} + \beta\,\mathrm{diag}(\mathbf{u}^2)$	1	ground state [21, 24]
a_0-flow	$-\Delta_h + \mathbb{V}$	2	critical point [11]
H^1-flow	$-\Delta_h + \alpha\mathbb{I}$	2	critical point [12]
DYS I	$\mathbb{I} - \gamma\Delta_h$	**1**	**ground state** (Theorem 3.4)

TABLE 1. Elliptic solvers needed by one iteration, and the strongest convergence result proved for each method: *ground state* and *critical point* mean global convergence, levels (iii) and (ii) of §1.2.

1.4. **Contributions of the present work.** The contributions of the present work are twofold. First, we propose the two efficient splitting schemes DYS I and DYS II of (1.6) and (1.7), whose single solve per iteration involves only $\mathbb{I} - \gamma\Delta_h$ or, with $\mathbb{V}_1$ the separable part of the potential, $\mathbb{I} - \gamma\Delta_h + \gamma\mathbb{V}_1$. Second, we prove global convergence to the unique positive discrete ground state for (1.6), to our knowledge the first such result for a method that inverts only a shifted Laplacian. Our main result (Theorem 3.4) states that, if using a monotone discrete Laplacian, then for every positive normalized initial guess $\mathbf{u}^0 > 0$ and every step size $0 < \gamma \le \gamma_{\max}$, with $\gamma_{\max} > 0$ explicitly computable from the data, the iteration (1.6) converges to the unique positive discrete ground state. Once the iterates are close to the ground state, the convergence is linear (Theorem 5.1). For simplicity we prove the theorem only for DYS I, but all arguments apply to DYS II, since $\mathbb{I} - \gamma\Delta_h + \gamma\mathbb{V}_1$ differs from $\mathbb{I} - \gamma\Delta_h$ by the nonnegative diagonal matrix $\gamma\mathbb{V}_1$, which preserves the M-matrix structure of the negative discrete Laplacian and the entrywise positive inverse used in the analysis, see Remark 3.6. Alongside the theorem, Section 6 compares (1.6) and its variant (1.7) with five efficient methods on three-dimensional problems with up to 999^3 unknowns on one GPU: with a simple step size rule, the Davis–Yin schemes match the conjugate gradient methods in wall-clock time and are much more robust with respect to the choice of the initial guess.

The general convergence theory for the Davis–Yin splitting requires all three parts of the objective to be convex [17], so the convergence of (1.6) has to be established from scratch due to the unit sphere constraint. The proof of Theorem 3.4 overcomes two difficulties. First, the energy of DYS I is not guaranteed to decrease monotonically, thus instead we analyze the radial direction as well as the tangential one, through a Lyapunov function that couples the energy of the normalized iterate with a radial residual $R^k = \|\mathbf{z}^k\|_2 - 1 + \gamma\lambda(\mathbf{u}^k)$, where $\lambda(\mathbf{u}^k)$ is the discrete eigenvalue functional of (3.3). Second, one has to prove that the iterates stay positive, and the proof is significantly different from and much more difficult than those for the a_u-flow and the GFDN, since one step of (1.6) reads $\mathbf{z}^{k+1} = (\|\mathbf{z}^k\|_2 - 1)\mathbf{u}^k + \mathbf{x}^k$, and the first term is negative whenever $\|\mathbf{z}^k\|_2 < 1$.

1.5. **Organization of the paper.** Section 2 presents the discrete ground state problem and the structural properties used in the paper. Section 3 introduces the quantities in which the analysis is carried out, states the main theorem, and describes the three steps of our proof strategy. Section 4 derives two exact identities for one step, proves positivity, the radial interval and a conditional descent estimate for the Lyapunov function, and removes the condition by an energy barrier argument. In Section 5, we prove the global convergence to the discrete ground state, and also prove a local linear rate. Numerical tests are given in Section 6, and concluding remarks are in Section 7.

2. The discrete ground state problem

We discretize (1.3) in space by the classical continuous finite element method with quadrature, following [12], using either Q^1 elements with Gauss–Lobatto quadrature on a uniform rectangular mesh, which coincides with the classical second-order finite difference scheme, or P^1 elements with mass lumping on a simplicial mesh satisfying a suitable angle condition. See [12] for their construction, the precise mesh condition, and the monotonicity properties that follow, which are also briefly reviewed in Appendix A.

Let $\boldsymbol{x}_i$ $(i = 1, \ldots, N)$ be the interior nodes of the mesh, let $w_i > 0$ be the quadrature weight (in either scheme above) at $\boldsymbol{x}_i$, let ϕ_i be the nodal basis functions, and let $\langle \cdot, \cdot \rangle$ denote the quadrature approximation of the L^2 inner product. A function is represented by its nodal values $\mathbf{u} = \begin{bmatrix} u_1 & \cdots & u_N \end{bmatrix}^\top \in \mathbb{R}^N$. Set

$$\mathbb{M} = \operatorname{diag}\{w_1, \cdots, w_N\}, \qquad \mathbb{V} = \operatorname{diag}\{V_1, \cdots, V_N\} \text{ with } V_i = V(\boldsymbol{x}_i), \qquad \mathbb{S}_{ij} = \langle \nabla \phi_i, \nabla \phi_j \rangle,$$

so that $\mathbb{M}$ is the lumped mass matrix and $\mathbb{S}$ the stiffness matrix. For a symmetric matrix A we write $A \succeq 0$ for positive semidefiniteness and $A \preceq B$ for $B - A \succeq 0$. The discrete counterpart of (1.4) is the nonlinear eigenvalue problem

$$\mathbb{S}\mathbf{u} + \mathbb{M}\mathbb{V}\mathbf{u} + \beta \mathbb{M}\mathbf{u}^3 = \lambda_h \mathbb{M}\mathbf{u}, \tag{2.1}$$

where $\mathbf{u}^3 := \begin{bmatrix} u_1^3 & \cdots & u_N^3 \end{bmatrix}^\top$. Equivalently, with the discrete Laplacian $\Delta_h := -\mathbb{M}^{-1}\mathbb{S}$,

$$-\Delta_h \mathbf{u} + \mathbb{V}\mathbf{u} + \beta \mathbf{u}^3 = \lambda_h \mathbf{u}. \tag{2.2}$$

For $\mathbf{v} \in \mathbb{R}^N$ set $\|\mathbf{v}\|_p^p = \sum_i w_i |v_i|^p$ for $1 \le p < \infty$ and $\|\mathbf{v}\|_\infty = \max_i |v_i|$. For the case $p = 2$, we denote the discrete L^2 inner product and norm by

$$\langle \mathbf{u}, \mathbf{v} \rangle_h := \mathbf{u}^\top \mathbb{M} \mathbf{v}, \qquad \|\mathbf{u}\|_2 := \langle \mathbf{u}, \mathbf{u} \rangle_h^{1/2}. \tag{2.3}$$

Thus the discrete energy is

$$\mathbf{E}_h(\mathbf{u}) = \frac{1}{2} \langle -\Delta_h \mathbf{u}, \mathbf{u} \rangle_h + \frac{1}{2} \langle \mathbb{V}\mathbf{u}, \mathbf{u} \rangle_h + \frac{\beta}{4} \|\mathbf{u}\|_4^4, \tag{2.4}$$

and the *discrete ground state problem* is

$$\min_{\mathbf{u} \in \mathbb{R}^N} \mathbf{E}_h(\mathbf{u}) \qquad \text{subject to} \qquad \|\mathbf{u}\|_2 = 1. \tag{2.5}$$

The constraint set is the manifold $\mathcal{M} := \{\mathbf{u} \in \mathbb{R}^N : \|\mathbf{u}\|_2 = 1\}$, the unit sphere of the discrete L^2 norm. Any minimizer of (2.5) solves (2.1), and taking the $\langle \cdot, \cdot \rangle_h$-product of (2.2) with $\mathbf{u}$ gives, as in (1.5), $\lambda_h = 2\mathbf{E}_h(\mathbf{u}) + \frac{\beta}{2} \|\mathbf{u}\|_4^4$. At a fixed vector $\mathbf{u}$, the linearized operator of (2.2) is

$$A_{\mathbf{u}} = -\Delta_h + \mathbb{V} + \beta \operatorname{diag}(\mathbf{u}^2). \tag{2.6}$$

We assume that the spatial discretization is monotone:

Assumption 2.1 (Monotone conforming discretization). *Let $\Omega \subset \mathbb{R}^d$, $d \le 3$, be a bounded polytope on which elliptic regularity holds, discretized by a shape-regular mesh, and let $V_i \ge 0$ for every i and $\beta > 0$. The spatial discretization is a monotone scheme constructed from a conforming finite element method for the Laplacian, that is, one of the two schemes of Appendix A, in both cases with the quadrature described there, so that the mass matrix is diagonal:*

(a) *the Q^1 finite element method on a uniform rectangular mesh, with the two-point Gauss–Lobatto rule in each variable on each cell, which coincides with the classical second-order finite difference scheme, or*
(b) *the P^1 finite element method with the vertex rule on a simplicial mesh satisfying the angle condition* (A.4)*, which in two dimensions a Delaunay triangulation satisfies.*

The properties of these schemes that the analysis uses are collected in the following lemma, and no others are used.

Lemma 2.2 (What the discretization provides). *Let Assumption 2.1 hold. Then the following hold.*

(i) *The mass matrix is diagonal, $\mathbb{M} = \operatorname{diag}\{w_1, \dots, w_N\}$, with $w_i > 0$ for every i.*
(ii) *The stiffness matrix satisfies $\mathbb{S} = \mathbb{S}^\top \succeq 0$ and $\mathbb{S}_{ij} \le 0$ for every $i \ne j$, and for every $\gamma > 0$ the inverse $(\mathbb{M} + \gamma\mathbb{S})^{-1}$ has strictly positive entries.*
(iii) *There is a constant C_S, independent of the mesh size, such that, for any $\mathbf{v} \in \mathbb{R}^N$,*

$$\|\mathbf{v}\|_6 \ \le\ C_S \ \|\mathbf{v}\|_{1,h}\,, \qquad \|\mathbf{v}\|_{1,h}^2 := \langle \mathbf{v}, -\Delta_h \mathbf{v}\rangle_h + \|\mathbf{v}\|_2^2\,, \tag{2.7}$$

where $\|\cdot\|_{1,h}$ is the discrete H^1 norm.

Proof. Item (i) is the quadrature, and the two sign conditions in item (ii) are the monotonicity of the discrete Laplacian, established for both schemes in Appendix A. For scheme (b) the second is exactly what (A.4) provides [37, Lemma 2.1]. Given them, $\mathbb{M} + \gamma\mathbb{S}$ is symmetric and positive definite, because $\mathbf{v}^\top(\mathbb{M} + \gamma\mathbb{S})\mathbf{v} = \mathbf{v}^\top\mathbb{M}\mathbf{v} + \gamma\,\mathbf{v}^\top\mathbb{S}\mathbf{v} \ge \mathbf{v}^\top\mathbb{M}\mathbf{v} > 0$ for $\mathbf{v} \ne 0$. Its off-diagonal entries are $\gamma\mathbb{S}_{ij} \le 0$ for $i \ne j$, since $\mathbb{M}$ is diagonal, and its adjacency graph is connected. The inverse of an irreducible Stieltjes matrix is entrywise positive [7, Chapter 6], which is the last assertion of item (ii). Item (iii) is [12, Lemma 5.5(iv)] together with the norm equivalences of its appendix, which identify $\|\cdot\|_{1,h}$ with the H^1 norm and the quadrature norms with the Lebesgue norms, all with constants independent of the mesh size. The elliptic regularity of Ω required in Assumption 2.1 enters only here, as the hypothesis under which the equivalence between $\|\cdot\|_{1,h}$ and the H^1 norm is proven in the appendix of [12]. Conformity of the two schemes and $d \le 3$ are what make the continuous embedding $H^1 \hookrightarrow L^6$ available. □

Remark 2.3 (The discrete ground state). Under Assumption 2.1 the matrix $\mathbb{M}A_{\mathbf{u}} = \mathbb{S} + \mathbb{M}\mathbb{V} + \beta\mathbb{M}\operatorname{diag}(\mathbf{u}^2)$ is an irreducible symmetric M-matrix for every $\mathbf{u}$. As a consequence, (2.5) has a minimizer that is unique up to sign, whose positive representative $\mathbf{u}_{\mathrm{GS}}$ is called the *discrete ground state*. It is the unique nonnegative minimizer of $\mathbf{E}_h$ on $\mathcal{M}$, and every nonnegative critical point of $\mathbf{E}_h|_{\mathcal{M}}$ coincides with it. See Appendix A for the argument and for references.

3. The scheme, the Lyapunov function, and the main theorem

The symbols $\succeq$ and $\preceq$ are used only for symmetric matrices such as $\mathbb{S}$ and $\mathbb{M} + \gamma\mathbb{S}$. Most operators of this paper need not be symmetric matrices, among them the discrete Laplacian $-\Delta_h = \mathbb{M}^{-1}\mathbb{S}$, the linearized operator $A_{\mathbf{u}}$, the tangent projection $\mathcal{P}_{\mathbf{u}}$ and the resolvent $\mathbb{G}_X^{-1}$ introduced below. Each of them is instead *h-self-adjoint*, meaning that $\langle \mathbf{u}, A\mathbf{v}\rangle_h = \langle A\mathbf{u}, \mathbf{v}\rangle_h$ for all $\mathbf{u}, \mathbf{v}$, where $\langle\cdot,\cdot\rangle_h$ is the discrete inner product (2.3). Since $\langle \mathbf{u}, A\mathbf{v}\rangle_h = \mathbf{u}^\top \mathbb{M} A\mathbf{v}$, this holds exactly when the matrix $\mathbb{M}A$ is symmetric, which implies that such an operator has real eigenvalues and an $\langle\cdot,\cdot\rangle_h$-orthonormal eigenbasis. Inequalities between h-self-adjoint operators are always written out as inequalities between the associated quadratic forms, never with $\preceq$.

3.1. **Norms, energy, and gradients.** The tangent projection at $\mathbf{u} \in \mathcal{M}$ is $\mathcal{P}_{\mathbf{u}} := \mathbb{I} - \mathbf{u}\mathbf{u}^\top\mathbb{M}$. Writing $w_{\min} := \min_i w_i > 0$, one has the pointwise bound

$$\|\mathbf{v}\|_\infty \le w_{\min}^{-1/2}\,\|\mathbf{v}\|_2 \qquad (\mathbf{v} \in \mathbb{R}^N), \tag{3.1}$$

since $w_i|v_i|^2 \le \|\mathbf{v}\|_2^2$ for every i. The bound holds only since the mesh is fixed and $\mathbb{R}^N$ is finite dimensional, and there is no such bound in the continuum. Its constant is of order $h^{-d/2}$ on quasi-uniform meshes, and it is the weakest of the three pointwise bounds collected in Lemma 3.2(ii). Those bounds and the stiffness-to-mass ratio Λ_{d} are the only place where the mesh enters the step-size threshold, see Remark 3.3.

The gradient of the discrete energy (2.4) with respect to $\langle\cdot,\cdot\rangle_h$ is

$$\nabla\mathbf{E}_h(\mathbf{u}) = -\Delta_h\mathbf{u} + \mathbb{V}\mathbf{u} + \beta\mathbf{u}^3. \tag{3.2}$$

The linearized operator (2.6) is $A_{\mathbf{u}} = -\Delta_h + \mathbb{V} + \beta\,\mathrm{diag}(\mathbf{u}^2)$, and $\nabla\mathbf{E}_h(\mathbf{u}) = A_{\mathbf{u}}\mathbf{u}$. The eigenvalue functional and the tangential part of the gradient are

$$\lambda(\mathbf{u}) := \langle \nabla\mathbf{E}_h(\mathbf{u}), \mathbf{u}\rangle_h, \qquad \mathbf{g}(\mathbf{u}) := \mathcal{P}_{\mathbf{u}}\nabla\mathbf{E}_h(\mathbf{u}) = \nabla\mathbf{E}_h(\mathbf{u}) - \lambda(\mathbf{u})\,\mathbf{u}. \tag{3.3}$$

A point $\mathbf{u} \in \mathcal{M}$ is a critical point of $\mathbf{E}_h|_{\mathcal{M}}$ when $\mathbf{g}(\mathbf{u}) = 0$, equivalently $A_{\mathbf{u}}\mathbf{u} = \lambda(\mathbf{u})\mathbf{u}$.

By (3.3) the vector $\mathbf{g}(\mathbf{u})$ is the projection of the gradient $\nabla\mathbf{E}_h(\mathbf{u})$ onto the tangent plane of $\mathcal{M}$ at $\mathbf{u}$, taken in the discrete L^2 metric $\langle\cdot,\cdot\rangle_h$. We call it the *projected gradient*. It coincides with the Riemannian gradient of $\mathbf{E}_h$ on $\mathcal{M}$ for that metric, written $\nabla_h^{\mathcal{R}}\mathbf{E}_h(\mathbf{u})$ in [12], but we avoid that name here, since the scheme (1.6) is not a Riemannian gradient method. In contrast, the H^1 flow is the Riemannian gradient descent method using the H^1 metric but needs to invert the shifted Laplacian twice per iteration [12].

3.2. **The scheme, the radial residual, and the Lyapunov function.** The metric matrix of the scheme is $\mathbb{G}_X := \mathbb{I} - \gamma\Delta_h$, so that $\mathbb{G}_X^{-1} = (\mathbb{M} + \gamma\mathbb{S})^{-1}\mathbb{M}$ is the resolvent appearing in the iteration. The associated X-inner product and X-norm are

$$\langle \mathbf{v}, \mathbf{w}\rangle_X := \langle \mathbf{v}, \mathbb{G}_X^{-1}\mathbf{w}\rangle_h, \qquad \|\mathbf{v}\|_X := \langle \mathbf{v}, \mathbf{v}\rangle_X^{1/2}. \tag{3.4}$$

The operator $\mathbb{G}_X^{-1}$ need not be a symmetric matrix, but it is h-self-adjoint in the sense of §3, since $\mathbb{M}\mathbb{G}_X^{-1} = \mathbb{M}(\mathbb{M} + \gamma\mathbb{S})^{-1}\mathbb{M}$ is symmetric. All properties of the X-norm used below follow from the spectral decomposition of $\mathbb{G}_X^{-1}$. Let $\mu_j \ge 0$ be the eigenvalues of $-\Delta_h$ for an $\langle\cdot,\cdot\rangle_h$-orthonormal eigenbasis, in which $\mathbb{G}_X^{-1}$ has the eigenvalues $1/(1+\gamma\mu_j) \in (0,1]$, and let $\hat{v}_j$ be the coefficients of $\mathbf{v}$ in that basis. Then

$$\|\mathbf{v}\|_X^2 = \sum_j \frac{\hat{v}_j^2}{1+\gamma\mu_j}, \qquad \left\|\mathbb{G}_X^{-1}\mathbf{v}\right\|_2^2 = \sum_j \frac{\hat{v}_j^2}{(1+\gamma\mu_j)^2},$$

and $0 < (1+\gamma\mu_j)^{-2} \le (1+\gamma\mu_j)^{-1} \le 1$ and $\frac{\gamma\mu_j}{1+\gamma\mu_j} \le \gamma\mu_j$ imply

$$\mathbf{v} \neq 0 \implies 0 < \|\mathbf{v}\|_X \le \|\mathbf{v}\|_2\,, \qquad \left\|\mathbb{G}_X^{-1}\mathbf{v}\right\|_2 \le \|\mathbf{v}\|_X\,, \tag{3.5}$$

$$\mathbf{u} \in \mathcal{M} \implies 0 \le 1 - \|\mathbf{u}\|_X^2 = \sum_j \frac{\gamma\mu_j}{1+\gamma\mu_j}\hat{u}_j^2 \le \gamma\langle \mathbf{u}, -\Delta_h\mathbf{u}\rangle_h. \tag{3.6}$$

In particular $\langle\cdot,\cdot\rangle_X$ is an inner product. We also use the norm of the metric matrix itself,

$$\|\mathbf{v}\|_{\mathcal{G}}^2 := \langle \mathbf{v}, \mathbb{G}_X\mathbf{v}\rangle_h = \|\mathbf{v}\|_2^2 + \gamma\langle \mathbf{v}, -\Delta_h\mathbf{v}\rangle_h, \qquad \text{so that} \qquad \|\mathbf{v}\|_2 \le \|\mathbf{v}\|_{\mathcal{G}}\,, \quad \|\mathbb{G}_X\mathbf{v}\|_X = \|\mathbf{v}\|_{\mathcal{G}}\,. \tag{3.7}$$

The descent of one step is measured in $\|\cdot\|_{\mathcal{G}}$ in Section 4 and converted to $\|\cdot\|_X$ in Proposition 4.8.

One step of the scheme is (1.6) with the iteration index dropped: it maps $\mathbf{z} \neq 0$ to $\mathbf{u} = \mathbf{z}/\|\mathbf{z}\|_2$, $\mathbf{x}$, and $\mathbf{z}^+ = \mathbf{z} + \mathbf{x} - \mathbf{u}$. The iteration is $\mathbf{z}^{k+1} = (\mathbf{z}^k)^+$ started from the normalized positive initial vector $\mathbf{z}^0 = \mathbf{u}^0$ with $\mathbf{u}^0 > 0$ and $\left\|\mathbf{u}^0\right\|_2 = 1$. For every $\mathbf{z} \neq 0$, with $\mathbf{u}$ the normalized iterate of (1.6), define the *radial residual* and the *Lyapunov function*

$$R(\mathbf{z}) := \|\mathbf{z}\|_2 - 1 + \gamma\lambda(\mathbf{u}), \qquad \mathcal{L}(\mathbf{z}) := \mathbf{E}_h(\mathbf{u}) + \frac{R(\mathbf{z})^2}{\gamma}. \tag{3.8}$$

Note that $R = 0$ is equivalent to $\|\mathbf{z}\|_2 = 1 - \gamma\lambda(\mathbf{u})$, which is the length of a fixed point of (1.6), and R carries no information on the direction of $\mathbf{z}$. The deviation in the tangential directions is measured by $\mathbf{g}(\mathbf{u})$, and at a fixed point both R and $\mathbf{g}(\mathbf{u})$ vanish, see (4.4). Whenever $\mathbf{z}^k \neq 0$ we abbreviate $\mathbf{u}^k := \mathbf{z}^k/\left\|\mathbf{z}^k\right\|_2$, $R^k := R(\mathbf{z}^k)$ and $\mathcal{L}^k := \mathcal{L}(\mathbf{z}^k)$, together with $\mathbf{g}^k := \mathbf{g}(\mathbf{u}^k)$. The superscript $+$ marks the same quantities after one step, at the next iteration index, e.g., $\mathbf{u}^+ := \mathbf{z}^+/\|\mathbf{z}^+\|_2$ is $\mathbf{u}^{k+1}$ whenever $\mathbf{z}^+ \neq 0$, and $R^+ := R(\mathbf{z}^+)$.

3.3. **Data constants and the step-size threshold.** The energy $\mathbf{E}_h(\mathbf{u}^0)$ and the eigenvalue functional $\lambda(\mathbf{u}^0)$ of the initial vector enter through the level

$$E_0 := \mathbf{E}_h(\mathbf{u}^0) + \frac{\lambda(\mathbf{u}^0)^2}{\max\{1, 128\,\mathbf{E}_h(\mathbf{u}^0)\}}. \tag{3.9}$$

Since $\frac{\beta}{4}\|\mathbf{u}\|_4^4 \le \mathbf{E}_h(\mathbf{u})$, the identity $\lambda(\mathbf{u}) = 2\mathbf{E}_h(\mathbf{u}) + \frac{\beta}{2}\|\mathbf{u}\|_4^4$ implies $\lambda(\mathbf{u}^0) \le 4\mathbf{E}_h(\mathbf{u}^0)$. Since $\mathbf{E}_h(\mathbf{u}^0)/\max\{1, 128\,\mathbf{E}_h(\mathbf{u}^0)\} \le \frac{1}{128}$ in both cases of the maximum, we have

$$\frac{\lambda(\mathbf{u}^0)^2}{\max\{1, 128\,\mathbf{E}_h(\mathbf{u}^0)\}} \le \frac{16\,\mathbf{E}_h(\mathbf{u}^0)^2}{\max\{1, 128\,\mathbf{E}_h(\mathbf{u}^0)\}} \le \frac{16}{128}\,\mathbf{E}_h(\mathbf{u}^0) = \frac{1}{8}\,\mathbf{E}_h(\mathbf{u}^0),$$

and hence $\mathbf{E}_h(\mathbf{u}^0) \le E_0 \le \frac{9}{8}\mathbf{E}_h(\mathbf{u}^0)$. We will show that $\mathcal{L}(\mathbf{z}^k) \le E_0$ for every k. Let $\Lambda_{\mathrm{d}} := \max_i \mathbb{S}_{ii}/w_i$ be the stiffness-to-mass ratio of the discrete Laplacian. Note that $\Lambda_{\mathrm{d}} > 0$, since $\mathbb{S}_{ii}$ is the value of the quadratic form of $\mathbb{S}$ at the ith unit vector and $\mathbb{S}$ is positive definite, see Appendix A. With $w_{\min} = \min_i w_i$ as in §3.1, set

$$P_0 := C_S^2\big(2E_0 + 1\big), \qquad U_0 := \min\Big\{\frac{1}{w_{\min}},\ \frac{P_0}{w_{\min}^{1/3}},\ 2\sqrt{\frac{E_0}{\beta\,w_{\min}}}\Big\}, \tag{3.10}$$

$$B_0 := \|V\|_\infty + \beta\min\{P_0^{3/2}, U_0\}, \qquad D := E_0 + B_0, \qquad B_1 := \|V\|_\infty + \beta U_0, \tag{3.11}$$

$$C_0 := 3\beta C_S P_0, \qquad L_0 := \|V\|_\infty + \min\big\{3\beta U_0,\ C_0 + 2\max\{C_0 - \tfrac{1}{4}, 0\}^2\big\}. \tag{3.12}$$

Here P_0 bounds $\|\mathbf{u}\|_6^2$ and U_0 bounds $\|\mathbf{u}\|_\infty^2$, while B_0, B_1 and L_0 bound the explicit part

$$\mathbf{n}(\mathbf{u}) := \mathbb{V}\mathbf{u} + \beta\mathbf{u}^3, \qquad \nabla\mathbf{E}_h(\mathbf{u}) = -\Delta_h\mathbf{u} + \mathbf{n}(\mathbf{u}), \tag{3.13}$$

of the gradient and its Jacobian, for normalized vectors of energy at most E_0, see Lemmas 3.2 and 4.3. The step-size threshold of Theorem 3.4 is

$$\gamma_{\max} := \min\Big\{1,\ \frac{1}{128D},\ \frac{1}{8L_0},\ \frac{2}{B_1 + \sqrt{B_1^2 + 12D\Lambda_{\mathrm{d}}}}\Big\}. \tag{3.14}$$

Every entry is an explicit function of the data. The proof uses $\gamma_{\max}$ only through the four conditions of the next lemma.

Lemma 3.1 (The step-size conditions)**.** *A step size $\gamma > 0$ satisfies $\gamma \le \gamma_{\max}$ if and only if*

$$\text{(C1)}\ \ \gamma \le 1, \qquad \text{(C2)}\ \ \gamma D \le \tfrac{1}{128}, \qquad \text{(C3)}\ \ \gamma L_0 \le \tfrac{1}{8}, \qquad \text{(C4)}\ \ \gamma B_1 + 3\gamma^2 D\Lambda_{\mathrm{d}} \le 1. \tag{3.15}$$

Under these conditions $\gamma \le 1/\max\{1, 128\,\mathbf{E}_h(\mathbf{u}^0)\}$, and therefore $\mathcal{L}(\mathbf{z}^0) = \mathbf{E}_h(\mathbf{u}^0) + \gamma\lambda(\mathbf{u}^0)^2 \le E_0$ for $\mathbf{z}^0 = \mathbf{u}^0$.

Proof. (C1), (C2) and (C3) are the first three entries of (3.14). The quadratic $3D\Lambda_{\mathrm{d}}\gamma^2 + B_1\gamma - 1$ in (C4) has a positive leading coefficient and the value -1 at $\gamma = 0$, thus it has one negative and one positive root, and it is nonpositive for $\gamma > 0$ if and only if γ is at most the positive root

$$\frac{-B_1 + \sqrt{B_1^2 + 12D\Lambda_{\mathrm{d}}}}{6D\Lambda_{\mathrm{d}}} = \frac{2}{B_1 + \sqrt{B_1^2 + 12D\Lambda_{\mathrm{d}}}},$$

which is the last entry of (3.14). Since $D \ge E_0 \ge \mathbf{E}_h(\mathbf{u}^0)$, (C1) and (C2) imply that γ is at most $1/\max\{1, 128\,\mathbf{E}_h(\mathbf{u}^0)\}$. For $\mathbf{z}^0 = \mathbf{u}^0$ the residual is $R(\mathbf{z}^0) = \gamma\lambda(\mathbf{u}^0)$, which implies $\mathcal{L}(\mathbf{z}^0) = \mathbf{E}_h(\mathbf{u}^0) + \gamma\lambda(\mathbf{u}^0)^2 \le \mathbf{E}_h(\mathbf{u}^0) + \lambda(\mathbf{u}^0)^2/\max\{1, 128\,\mathbf{E}_h(\mathbf{u}^0)\} = E_0$. ☐

The next lemma collects the bounds implied by the energy level E_0, which are used throughout Section 4.

Lemma 3.2 (Consequences of the energy bound)**.** *Let $\mathbf{v} \in \mathcal{M}$ satisfy $\mathbf{E}_h(\mathbf{v}) \le E_0$. Then the following hold.*

(i) $\langle \mathbf{v}, -\Delta_h \mathbf{v}\rangle_h \le 2E_0$, $\|\mathbf{v}\|_{1,h}^2 \le 2E_0 + 1$ *and* $\|\mathbf{v}\|_6^2 \le P_0$.
(ii) $\|\mathbf{v}\|_\infty^2 \le U_0$.
(iii) $\|\mathbf{v}^3\|_2 \le \min\{P_0^{3/2}, U_0\}$, $\|\mathbf{v}\|_4^4 \le \min\{P_0^{3/2}, U_0\}$, $\|\mathbf{n}(\mathbf{v})\|_2 \le B_0$ *and* $\langle \mathbf{n}(\mathbf{v}), \mathbf{v}\rangle_h \le B_0$.
(iv) $0 \le \lambda(\mathbf{v}) \le 2D$, *and thus* $0 \le \gamma\lambda(\mathbf{v}) \le 2\gamma D$.

Proof. (i) Each of the three terms of $\mathbf{E}_h$ in (2.4) is nonnegative under Assumption 2.1, which implies

$$\tfrac{1}{2}\langle \mathbf{v}, -\Delta_h \mathbf{v}\rangle_h \le \mathbf{E}_h(\mathbf{v}) \le E_0, \qquad \|\mathbf{v}\|_{1,h}^2 = \langle \mathbf{v}, -\Delta_h \mathbf{v}\rangle_h + \|\mathbf{v}\|_2^2 \le 2E_0 + 1,$$

and the Sobolev embedding (2.7) implies $\|\mathbf{v}\|_6^2 \le C_S^2(2E_0 + 1) = P_0$.

(ii) A single term of a sum of nonnegative numbers is at most the whole sum. Applied to $\|\mathbf{v}\|_2^2 = \sum_i w_i v_i^2 = 1$, to $\|\mathbf{v}\|_6^6 = \sum_i w_i v_i^6 \le P_0^3$ and to $\frac{\beta}{4}\|\mathbf{v}\|_4^4 = \frac{\beta}{4}\sum_i w_i v_i^4 \le E_0$, this implies $v_i^2 \le w_i^{-1}$, $v_i^2 \le P_0 w_i^{-1/3}$ and $v_i^2 \le 2\sqrt{E_0/(\beta w_i)}$ for every i, and $w_i \ge w_{\min}$.

(iii) $\|\mathbf{v}^3\|_2 = \|\mathbf{v}\|_6^3 \le P_0^{3/2}$, and $\|\mathbf{v}^3\|_2 \le \|\mathbf{v}\|_\infty^2 \|\mathbf{v}\|_2 \le U_0$. The Cauchy–Schwarz inequality applied to the pair $(|\mathbf{v}|, |\mathbf{v}|^3)$ in $\langle\cdot,\cdot\rangle_h$ implies $\|\mathbf{v}\|_4^4 \le \|\mathbf{v}\|_2 \|\mathbf{v}^3\|_2$. Since $\|\mathbb{V}\mathbf{v}\|_2 \le \|V\|_\infty \|\mathbf{v}\|_2$ and $\langle \mathbb{V}\mathbf{v}, \mathbf{v}\rangle_h \le \|V\|_\infty$, both $\|\mathbf{n}(\mathbf{v})\|_2$ and $\langle \mathbf{n}(\mathbf{v}), \mathbf{v}\rangle_h = \langle \mathbb{V}\mathbf{v}, \mathbf{v}\rangle_h + \beta\|\mathbf{v}\|_4^4$ are at most $\|V\|_\infty + \beta\min\{P_0^{3/2}, U_0\} = B_0$.

(iv) $\lambda(\mathbf{v}) = \langle \mathbf{v}, -\Delta_h \mathbf{v}\rangle_h + \langle \mathbf{n}(\mathbf{v}), \mathbf{v}\rangle_h$ is a sum of nonnegative terms, and by (i) and (iii) it is at most $2E_0 + B_0 \le 2D$. ☐

Remark 3.3 (Mesh dependence and the size of the threshold). The mesh is fixed throughout and every constant above is a finite number determined by that mesh and the data, and no statement below involves a limit $h \to 0$. To see how the threshold behaves under refinement, consider a family of quasi-uniform meshes, so that $w_{\min}$ is of order h^d and Λ_{d} of order h^{-2}, and a family of initial vectors whose energies $\mathbf{E}_h(\mathbf{u}^0)$ are bounded independently of h, as are the normalized interpolants of a fixed smooth function that vanishes on $\partial\Omega$. Then the constants split into three groups. The first consists of universal numbers. The second consists of E_0, P_0, B_0, D and C_0, which are built from $\mathbf{E}_h(\mathbf{u}^0)$, from $\|V\|_\infty$ and β, and from the constant C_S of the Sobolev embedding (2.7), and are therefore bounded independently of h. The third consists of the pointwise bound U_0, of order at most $h^{-d/3}$ by its middle alternative, of B_1, which is built from it, and of Λ_{d}. Since $L_0 \le \|V\|_\infty + C_0 + 2\max\{C_0 - \frac{1}{4}, 0\}^2$ belongs to the second group, the mesh restricts $\gamma_{\max}$ only through the last entry of (3.14), which is of order h for $d \le 3$. The assumption on the initial energy cannot be dropped: the normalized constant vector used as the initial guess in Section 6 vanishes on the boundary, and this jump gives it an energy $\mathbf{E}_h(\mathbf{u}^0)$ of order h^{-1} on the finite difference grid, for which $\gamma_{\max}$ is of order h^2 in three dimensions. On the other hand, for a fixed mesh, potential and initial vector, the constants E_0, B_0, D, B_1 and L_0 grow at most linearly in β, thus $\gamma_{\max}$ is of order β^{-1} as $\beta \to \infty$, which is also the order of the step size rule used in Section 6.

3.4. **The main theorem.**

Theorem 3.4 (Global convergence to the discrete ground state). *Let Assumption 2.1 hold, start the iteration* (1.6) *from a positive normalized vector,* $\mathbf{z}^0 = \mathbf{u}^0 > 0$ *with* $\|\mathbf{u}^0\|_2 = 1$, *and let* $0 < \gamma \le \gamma_{\max}$. *Here* $\gamma_{\max} > 0$ *is a threshold, given in* (3.14), *determined by the initial vector* $\mathbf{u}^0$, *by the data* V *and* β, *by the constant* C_S *of the Sobolev embedding* (2.7), *by the mesh size through* $w_{\min} = \min_i w_i$, *and by the stiffness-to-mass ratio* Λ_{d} *of the discrete Laplacian. Then* $\mathbf{u}^k > 0$ *entrywise for every* $k \ge 0$, *and*

$$\mathbf{u}^k \to \mathbf{u}_{\mathrm{GS}}, \qquad \mathbf{E}_h(\mathbf{u}^k) \to \mathbf{E}_h(\mathbf{u}_{\mathrm{GS}}) = \min_{\mathcal{M}} \mathbf{E}_h,$$

where $\mathbf{u}_{\mathrm{GS}} > 0$ *is the positive global minimizer of* $\mathbf{E}_h$ *on* $\mathcal{M}$.

Remark 3.5 (The four step-size conditions). The four conditions of Lemma 3.1 enter the proof at separate places. (C2) makes the radial deviation $1 - \|\mathbf{z}^k\|_2$, the residual R^k and the change of direction in one step small, of order γD or $\sqrt{\gamma D}$. (C3) controls the second-order Taylor remainder of the explicit part of the energy, by whichever of the pointwise and the Sobolev estimates is better. (C4) is the positivity condition, and for the families of bounded initial energy in Remark 3.3 it is the only condition that restricts the step size through the mesh. (C1) is used only in Lemma 3.1, to place the initial value $\mathcal{L}^0$ below E_0. The descent inequality (4.31) of Proposition 4.8 has the universal coefficients $\frac{1}{4}$ and $\frac{1}{2}$.

Remark 3.6 (The variant DYS II). The analysis below applies without change to (1.7), with $V_1 \ge 0$ and $V_2 \ge 0$, after the substitutions

$$-\Delta_h \;\to\; -\Delta_h + \mathbb{V}_1, \qquad \mathbf{n}(\mathbf{u}) \;\to\; \mathbb{V}_2\mathbf{u} + \beta\mathbf{u}^3, \qquad \|V\|_\infty \;\to\; \|V_2\|_\infty, \qquad \Lambda_{\mathrm{d}} \;\to\; \max_i\big(\mathbb{S}_{ii}/w_i + V_{1,i}\big).$$

The energy and the level E_0 are unchanged, the metric matrix becomes $\mathbb{I} - \gamma\Delta_h + \gamma\mathbb{V}_1$, and the norms $\|\cdot\|_X$ and $\|\cdot\|_{\mathcal{G}}$ are built from it. The matrix $\mathbb{M} + \gamma\mathbb{S} + \gamma\mathbb{M}\mathbb{V}_1$ is again an irreducible Stieltjes matrix, hence Lemma 2.2(ii) holds for it, and since $V_1 \ge 0$ the quadratic form $\langle \mathbf{v}, (-\Delta_h + \mathbb{V}_1)\mathbf{v}\rangle_h$

is still at most $2\mathbf{E}_h(\mathbf{v})$ and still dominates $\langle \mathbf{v}, -\Delta_h \mathbf{v}\rangle_h$ in the Sobolev inequality (2.7). With these substitutions every constant of §3.3, every identity and every estimate below holds for DYS II.

Proof strategy. The proof consists of three steps. Steps 1 and 2 are in Section 4, with Step 2 in §4.5, and Step 3 is in Section 5.

Step 1 is the analysis of one step, from $\mathbf{z}^k$ to $\mathbf{z}^{k+1}$. Two exact identities express the change of energy and the new residual after one step in terms of the change of direction $\mathbf{u}^{k+1} - \mathbf{u}^k$, the two lengths $\|\mathbf{z}^k\|_2$ and $\|\mathbf{z}^{k+1}\|_2$, and the Taylor remainder of the explicit part of the energy. In the energy identity the implicit kinetic term of the step has a positive coefficient, and in the residual identity the leading kinetic variation of λ cancels. Positivity of $\mathbf{z}^{k+1}$ and the one-sided radial interval $1 - 3\gamma D \le \|\mathbf{z}^{k+1}\|_2 \le 1$ follow from the entrywise positive resolvent and the sign structure of the discretization, using only the energy bound of $\mathbf{u}^k$. If $\mathbf{u}^{k+1}$ also has energy at most E_0, the identities imply the conditional descent $\gamma(\mathcal{L}^k - \mathcal{L}^{k+1}) \ge \frac{5}{16}\|\mathbf{u}^{k+1} - \mathbf{u}^k\|_{\mathcal{G}}^2 + \frac{3}{4}(R^k)^2$.

Step 2 is an energy barrier argument. The condition on the new energy is removed by a continuation argument, in which the step size and the radial deviation of the input are scaled by the same factor. The scaled input has a smaller residual and a Lyapunov value still at most E_0, and a first contact of the energy of the new direction with the level E_0 would contradict the conditional descent. Hence the descent holds unconditionally, which is Proposition 4.8.

Step 3 is the propagation and the passage to the limit. A single induction propagates $\mathbf{u}^k > 0$, the radial interval and $\mathcal{L}^k \le E_0$. Summing the one-step drop along the whole sequence gives $\sum_k \gamma \|\mathbf{g}^k\|_X^2 < \infty$ and $\sum_k (R^k)^2/\gamma < \infty$, thus $\|\mathbf{g}^k\|_2 \to 0$ and $R^k \to 0$. Every accumulation point of $(\mathbf{u}^k)$ is therefore a critical point of $\mathbf{E}_h|_{\mathcal{M}}$, and it is nonnegative since it is a limit of positive iterates. Since the only nonnegative critical point is the positive global minimizer $\mathbf{u}_{\mathrm{GS}}$ (Remark 2.3), every accumulation point equals $\mathbf{u}_{\mathrm{GS}}$, and by compactness of $\mathcal{M}$ the whole sequence converges. This is Theorem 3.4. Once the iterates are close to $\mathbf{u}_{\mathrm{GS}}$, the same descent inequality implies a linear rate, which is Theorem 5.1.

Sections 4 and 5 are devoted to the proofs of Theorems 3.4 and 5.1.

4. Estimates and properties of one iteration

For convenience, in this section we drop the iteration index k and mark the next iteration by the superscript $+$: $\mathbf{z}$, $\mathbf{u}$ and $\mathbf{x}$ stand for $\mathbf{z}^k$, $\mathbf{u}^k$ and $\mathbf{x}^k$, and $\mathbf{z}^+$, $\mathbf{u}^+$, R^+ and $\mathcal{L}^+$ for $\mathbf{z}^{k+1}$, $\mathbf{u}^{k+1}$, R^{k+1} and $\mathcal{L}^{k+1}$, where $\mathbf{z} \ne 0$ and $\mathbf{u}^+ = \mathbf{z}^+/\|\mathbf{z}^+\|_2$ if $\mathbf{z}^+ \ne 0$. The change of direction is $\boldsymbol{\delta} := \mathbf{u}^+ - \mathbf{u}$. Since $\|\mathbf{u}\|_2 = \|\mathbf{u}^+\|_2 = 1$ and $\mathbf{u}^+ = \mathbf{z}^+/\|\mathbf{z}^+\|_2$, we have $\|\boldsymbol{\delta}\|_2^2 = \|\mathbf{u}^+\|_2^2 - 2\langle \mathbf{u}, \mathbf{u}^+\rangle_h + \|\mathbf{u}\|_2^2 = 2 - 2\langle \mathbf{u}, \mathbf{u}^+\rangle_h$ and $\langle \mathbf{u}, \boldsymbol{\delta}\rangle_h = \langle \mathbf{u}, \mathbf{u}^+\rangle_h - 1$, i.e.,

$$\langle \mathbf{u}, \boldsymbol{\delta}\rangle_h = -\tfrac{1}{2}\|\boldsymbol{\delta}\|_2^2, \qquad \|\boldsymbol{\delta}\|_2^2 = 2\big(1 - \langle \mathbf{u}, \mathbf{u}^+\rangle_h\big) = 2\Big(1 - \frac{\langle \mathbf{u}, \mathbf{z}^+\rangle_h}{\|\mathbf{z}^+\|_2}\Big). \tag{4.1}$$

Let $H(\mathbf{u}) := \frac{1}{2}\langle \mathbb{V}\mathbf{u}, \mathbf{u}\rangle_h + \frac{\beta}{4}\|\mathbf{u}\|_4^4$ be the explicit part of the energy, as in Appendix B, so that $\mathbf{E}_h(\mathbf{u}) = \frac{1}{2}\langle \mathbf{u}, -\Delta_h \mathbf{u}\rangle_h + H(\mathbf{u})$ and the gradient of H with respect to $\langle\cdot,\cdot\rangle_h$ is $\mathbf{n}$ of (3.13). The Taylor remainder of H along $\boldsymbol{\delta}$ is

$$\begin{gathered}\mathcal{T}(\mathbf{u}, \boldsymbol{\delta}) := H(\mathbf{u}^+) - H(\mathbf{u}) - \langle \mathbf{n}(\mathbf{u}), \boldsymbol{\delta}\rangle_h = \int_0^1 (1-\sigma)\,\langle \mathbf{n}'(\mathbf{u} + \sigma\boldsymbol{\delta})\boldsymbol{\delta}, \boldsymbol{\delta}\rangle_h \, d\sigma, \\ \mathbf{n}'(\mathbf{w}) := \mathbb{V} + 3\beta\,\mathrm{diag}(\mathbf{w}^2),\end{gathered} \tag{4.2}$$

by Taylor's formula with integral remainder for the polynomial $\sigma \mapsto H(\mathbf{u}+\sigma\boldsymbol{\delta})$, whose second derivative is $\langle \mathbf{n}'(\mathbf{u}+\sigma\boldsymbol{\delta})\boldsymbol{\delta},\boldsymbol{\delta}\rangle_h$. Since the kinetic part of $\mathbf{E}_h$ is quadratic, we have $\frac{1}{2}\langle \mathbf{u}^+, -\Delta_h\mathbf{u}^+\rangle_h = \frac{1}{2}\langle \mathbf{u}, -\Delta_h\mathbf{u}\rangle_h + \langle -\Delta_h\mathbf{u},\boldsymbol{\delta}\rangle_h + \frac{1}{2}\langle \boldsymbol{\delta}, -\Delta_h\boldsymbol{\delta}\rangle_h$, and adding (4.2), we get the Taylor expansion of $\mathbf{E}_h$,

$$\mathbf{E}_h(\mathbf{u}^+) - \mathbf{E}_h(\mathbf{u}) = \langle \nabla\mathbf{E}_h(\mathbf{u}),\boldsymbol{\delta}\rangle_h + \tfrac{1}{2}\langle \boldsymbol{\delta}, -\Delta_h\boldsymbol{\delta}\rangle_h + \mathcal{T}(\mathbf{u},\boldsymbol{\delta}), \tag{4.3}$$

where $\nabla\mathbf{E}_h(\mathbf{u}) = -\Delta_h\mathbf{u} + \mathbf{n}(\mathbf{u})$ by (3.13).

4.1. **Exact identities for one step.**

Lemma 4.1 (Step identities)**.** *Let* $\mathbf{z}\neq 0$. *Then*

$$\mathbb{G}_X\mathbf{z}^+ = \mathbf{u} - \gamma\mathbf{n}(\mathbf{u}) - \gamma\big(1-\|\mathbf{z}\|_2\big)(-\Delta_h\mathbf{u}), \qquad \mathbb{G}_X\big(\mathbf{z}^+ - \mathbf{z}\big) = -\big(R\,\mathbf{u} + \gamma\,\mathbf{g}(\mathbf{u})\big). \tag{4.4}$$

If in addition $\mathbf{z}^+\neq 0$, *then*

$$\left\|\mathbf{z}^+\right\|_2 \mathbb{G}_X\boldsymbol{\delta} = \big(1-\left\|\mathbf{z}^+\right\|_2\big)\mathbf{u} - \gamma\nabla\mathbf{E}_h(\mathbf{u}) - \gamma\big(\left\|\mathbf{z}^+\right\|_2 - \|\mathbf{z}\|_2\big)(-\Delta_h\mathbf{u}), \tag{4.5}$$

$$\big(1+\gamma\langle \mathbf{u}, -\Delta_h\mathbf{u}\rangle_h\big)\big(\left\|\mathbf{z}^+\right\|_2 - \|\mathbf{z}\|_2\big) = -R - \left\|\mathbf{z}^+\right\|_2\Big(\gamma\langle -\Delta_h\mathbf{u},\boldsymbol{\delta}\rangle_h - \tfrac{1}{2}\left\|\boldsymbol{\delta}\right\|_2^2\Big). \tag{4.6}$$

Proof. By (1.6), we have $\mathbf{z}^+ - \mathbf{z} = \mathbf{x} - \mathbf{u}$ and $\mathbb{G}_X\mathbf{x} = 2\mathbf{u} - \mathbf{z} - \gamma\mathbf{n}(\mathbf{u})$, which imply

$$\mathbb{G}_X\mathbf{z}^+ = \mathbb{G}_X\mathbf{z} + 2\mathbf{u} - \mathbf{z} - \gamma\mathbf{n}(\mathbf{u}) - \mathbb{G}_X\mathbf{u} = (\mathbb{G}_X - \mathbb{I})(\mathbf{z}-\mathbf{u}) + \mathbf{u} - \gamma\mathbf{n}(\mathbf{u}),$$

and, since $\mathbb{G}_X - \mathbb{I} = -\gamma\Delta_h$ and $\mathbf{z}-\mathbf{u} = (\|\mathbf{z}\|_2 - 1)\mathbf{u}$, we have $(\mathbb{G}_X - \mathbb{I})(\mathbf{z}-\mathbf{u}) = (\|\mathbf{z}\|_2 - 1)(-\gamma\Delta_h)\mathbf{u}$, which implies the first identity. Since $\mathbf{z} = \|\mathbf{z}\|_2\,\mathbf{u}$, we have $\mathbb{G}_X\mathbf{z} = \|\mathbf{z}\|_2\,\mathbb{G}_X\mathbf{u} = \|\mathbf{z}\|_2\,\mathbf{u} + \gamma\,\|\mathbf{z}\|_2\,(-\Delta_h\mathbf{u})$, and subtracting it from the first identity, we have

$$\mathbb{G}_X(\mathbf{z}^+ - \mathbf{z}) = (1-\|\mathbf{z}\|_2)\mathbf{u} - \gamma(-\Delta_h\mathbf{u}) - \gamma\mathbf{n}(\mathbf{u}) = (1-\|\mathbf{z}\|_2)\mathbf{u} - \gamma\nabla\mathbf{E}_h(\mathbf{u}), \tag{4.7}$$

and $\nabla\mathbf{E}_h(\mathbf{u}) = \lambda(\mathbf{u})\mathbf{u} + \mathbf{g}(\mathbf{u})$ by (3.3) and the definition (3.8) of R imply the second identity. For (4.5), we have

$$\mathbf{z}^+ - \mathbf{z} = \left\|\mathbf{z}^+\right\|_2\mathbf{u}^+ - \|\mathbf{z}\|_2\,\mathbf{u} = \left\|\mathbf{z}^+\right\|_2\boldsymbol{\delta} + (\left\|\mathbf{z}^+\right\|_2 - \|\mathbf{z}\|_2)\mathbf{u}. \tag{4.8}$$

Applying $\mathbb{G}_X$ to (4.8), with $\mathbb{G}_X\mathbf{u} = \mathbf{u} + \gamma(-\Delta_h\mathbf{u})$, and using (4.7), we have

$$\left\|\mathbf{z}^+\right\|_2\mathbb{G}_X\boldsymbol{\delta} + (\left\|\mathbf{z}^+\right\|_2 - \|\mathbf{z}\|_2)\big(\mathbf{u} + \gamma(-\Delta_h\mathbf{u})\big) = \mathbb{G}_X(\mathbf{z}^+ - \mathbf{z}) = (1-\|\mathbf{z}\|_2)\mathbf{u} - \gamma\nabla\mathbf{E}_h(\mathbf{u}),$$

and moving the second term of the left side to the right side, with $(1-\|\mathbf{z}\|_2) - (\|\mathbf{z}^+\|_2 - \|\mathbf{z}\|_2) = 1 - \|\mathbf{z}^+\|_2$, we get (4.5). For (4.6), we take the $\langle\cdot,\cdot\rangle_h$-product of the second identity of (4.4) with $\mathbf{u}$ and use $\langle \mathbf{g}(\mathbf{u}),\mathbf{u}\rangle_h = 0$, the decomposition (4.8), $\langle \mathbb{G}_X\boldsymbol{\delta},\mathbf{u}\rangle_h = \langle \mathbf{u},\boldsymbol{\delta}\rangle_h + \gamma\langle -\Delta_h\mathbf{u},\boldsymbol{\delta}\rangle_h$, (4.1) and $\langle \mathbf{u}, \mathbb{G}_X\mathbf{u}\rangle_h = 1 + \gamma\langle \mathbf{u}, -\Delta_h\mathbf{u}\rangle_h$. □

Lemma 4.2 (Energy and residual identities)**.** *Let* $\mathbf{z}\neq 0$ *with* $\mathbf{z}^+\neq 0$. *Then*

$$\begin{aligned}\gamma\big(\mathbf{E}_h(\mathbf{u}) - \mathbf{E}_h(\mathbf{u}^+)\big) = {}& \frac{\|\mathbf{z}^+\|_2 + 1}{2}\,\|\boldsymbol{\delta}\|_2^2 + \gamma\Big(\left\|\mathbf{z}^+\right\|_2 - \frac{1}{2}\Big)\langle \boldsymbol{\delta}, -\Delta_h\boldsymbol{\delta}\rangle_h \\ & + \gamma\big(\left\|\mathbf{z}^+\right\|_2 - \|\mathbf{z}\|_2\big)\langle -\Delta_h\mathbf{u},\boldsymbol{\delta}\rangle_h - \gamma\,\mathcal{T}(\mathbf{u},\boldsymbol{\delta}),\end{aligned} \tag{4.9}$$

$$\begin{aligned}R^+ = {}& -\tfrac{1}{2}\,\|\boldsymbol{\delta}\|_2^2 - \gamma\big(\left\|\mathbf{z}^+\right\|_2 - \|\mathbf{z}\|_2\big)\langle -\Delta_h\mathbf{u},\mathbf{u}^+\rangle_h \\ & - \gamma\big(\left\|\mathbf{z}^+\right\|_2 - 1\big)\langle -\Delta_h\boldsymbol{\delta},\mathbf{u}^+\rangle_h + \gamma\langle \mathbf{n}(\mathbf{u}^+) - \mathbf{n}(\mathbf{u}),\mathbf{u}^+\rangle_h.\end{aligned} \tag{4.10}$$

Proof. Taking the $\langle\cdot,\cdot\rangle_h$-product of (4.5) with $\boldsymbol{\delta}$ and using (4.1), we have

$$\left\|\mathbf{z}^+\right\|_2\left(\|\boldsymbol{\delta}\|_2^2+\gamma\langle\boldsymbol{\delta},-\Delta_h\boldsymbol{\delta}\rangle_h\right)=-\frac{1-\|\mathbf{z}^+\|_2}{2}\|\boldsymbol{\delta}\|_2^2-\gamma\langle\nabla\mathbf{E}_h(\mathbf{u}),\boldsymbol{\delta}\rangle_h-\gamma\big(\left\|\mathbf{z}^+\right\|_2-\|\mathbf{z}\|_2\big)\langle-\Delta_h\mathbf{u},\boldsymbol{\delta}\rangle_h.$$

Eliminating $\gamma\langle\nabla\mathbf{E}_h(\mathbf{u}),\boldsymbol{\delta}\rangle_h$ between the equation above and (4.3) multiplied by γ, we obtain (4.9). For (4.10), we take the $\langle\cdot,\cdot\rangle_h$-product of the first identity of (4.4) with $\mathbf{u}^+$. Since $\mathbf{z}^+=\|\mathbf{z}^+\|_2\,\mathbf{u}^+$, we have $\langle\mathbb{G}_X\mathbf{z}^+,\mathbf{u}^+\rangle_h=\|\mathbf{z}^+\|_2\,\langle\mathbb{G}_X\mathbf{u}^+,\mathbf{u}^+\rangle_h=\|\mathbf{z}^+\|_2+\|\mathbf{z}^+\|_2\,\gamma\langle\mathbf{u}^+,-\Delta_h\mathbf{u}^+\rangle_h$, while $\langle\mathbf{u},\mathbf{u}^+\rangle_h=1-\frac{1}{2}\|\boldsymbol{\delta}\|_2^2$, thus

$$\left\|\mathbf{z}^+\right\|_2-1=-\tfrac{1}{2}\|\boldsymbol{\delta}\|_2^2+\gamma\big(\|\mathbf{z}\|_2-1\big)\langle-\Delta_h\mathbf{u},\mathbf{u}^+\rangle_h-\gamma\langle\mathbf{n}(\mathbf{u}),\mathbf{u}^+\rangle_h-\left\|\mathbf{z}^+\right\|_2\gamma\langle\mathbf{u}^+,-\Delta_h\mathbf{u}^+\rangle_h.$$

Adding $\gamma\lambda(\mathbf{u}^+)=\gamma\langle\mathbf{u}^+,-\Delta_h\mathbf{u}^+\rangle_h+\gamma\langle\mathbf{n}(\mathbf{u}^+),\mathbf{u}^+\rangle_h$, which follows from (3.3) and (3.13), to both sides, and writing $\|\mathbf{z}\|_2-1=(\|\mathbf{z}^+\|_2-1)-(\|\mathbf{z}^+\|_2-\|\mathbf{z}\|_2)$ and $\langle-\Delta_h\mathbf{u}^+,\mathbf{u}^+\rangle_h=\langle-\Delta_h\mathbf{u},\mathbf{u}^+\rangle_h+\langle-\Delta_h\boldsymbol{\delta},\mathbf{u}^+\rangle_h$, we get (4.10). □

4.2. **Estimates for the explicit part.**

Lemma 4.3 (The explicit part on an energy sublevel)**.** *Let (C3) in* (3.15) *hold, and assume* $\mathbf{z}\neq 0$*,* $\mathbf{z}^+\neq 0$*,* $\mathbf{E}_h(\mathbf{u})\le E_0$ *and* $\mathbf{E}_h(\mathbf{u}^+)\le E_0$*. Then*

$$\gamma\,\mathcal{T}(\mathbf{u},\boldsymbol{\delta})\le\tfrac{1}{16}\|\boldsymbol{\delta}\|_{\mathcal{G}}^2,\qquad\gamma\big|\langle\mathbf{n}(\mathbf{u}^+)-\mathbf{n}(\mathbf{u}),\mathbf{u}^+\rangle_h\big|\le 3\gamma B_0\|\boldsymbol{\delta}\|_2\le 3\gamma D\,\|\boldsymbol{\delta}\|_{\mathcal{G}}\,. \tag{4.11}$$

Proof. For $0\le\sigma\le 1$, the triangle inequality applied to $\mathbf{u}+\sigma\boldsymbol{\delta}=(1-\sigma)\mathbf{u}+\sigma\mathbf{u}^+$ and Lemma 3.2 at $\mathbf{u}$ and $\mathbf{u}^+$ give $\|\mathbf{u}+\sigma\boldsymbol{\delta}\|_\infty\le(1-\sigma)\|\mathbf{u}\|_\infty+\sigma\|\mathbf{u}^+\|_\infty\le\sqrt{U_0}$ and, in the same way, $\|\mathbf{u}+\sigma\boldsymbol{\delta}\|_6\le\sqrt{P_0}$. By the definition of $\mathbf{n}'$ in (4.2), it follows that

$$\langle\mathbf{n}'(\mathbf{u}+\sigma\boldsymbol{\delta})\boldsymbol{\delta},\boldsymbol{\delta}\rangle_h=\sum_i w_iV_i\delta_i^2+3\beta\sum_i w_i(u_i+\sigma\delta_i)^2\delta_i^2\le\|V\|_\infty\|\boldsymbol{\delta}\|_2^2+3\beta\sum_i w_i(u_i+\sigma\delta_i)^2\delta_i^2.$$

The last sum is bounded in two ways. First, $\sum_i w_i(u_i+\sigma\delta_i)^2\delta_i^2\le U_0\|\boldsymbol{\delta}\|_2^2$, which gives

$$\langle\mathbf{n}'(\mathbf{u}+\sigma\boldsymbol{\delta})\boldsymbol{\delta},\boldsymbol{\delta}\rangle_h\le(\|V\|_\infty+3\beta U_0)\|\boldsymbol{\delta}\|_2^2\,. \tag{4.12}$$

Second, Hölder's inequality with exponents 3 and $\frac{3}{2}$ gives $\sum_i w_i(u_i+\sigma\delta_i)^2\delta_i^2\le\|\mathbf{u}+\sigma\boldsymbol{\delta}\|_6^2\|\boldsymbol{\delta}\|_3^2\le P_0\|\boldsymbol{\delta}\|_3^2$, and Hölder's inequality with exponents $\frac{4}{3}$ and 4 gives

$$\|\boldsymbol{\delta}\|_3^3=\sum_i w_i|\delta_i|^{3/2}|\delta_i|^{3/2}\le\|\boldsymbol{\delta}\|_2^{3/2}\|\boldsymbol{\delta}\|_6^{3/2},\qquad\text{i.e.,}\qquad\|\boldsymbol{\delta}\|_3^2\le\|\boldsymbol{\delta}\|_2\|\boldsymbol{\delta}\|_6\,.$$

With the Sobolev embedding (2.7), $\|\boldsymbol{\delta}\|_6\le C_S\|\boldsymbol{\delta}\|_{1,h}$, we get

$$\begin{aligned}\langle\mathbf{n}'(\mathbf{u}+\sigma\boldsymbol{\delta})\boldsymbol{\delta},\boldsymbol{\delta}\rangle_h&\le\|V\|_\infty\|\boldsymbol{\delta}\|_2^2+C_0\|\boldsymbol{\delta}\|_2\|\boldsymbol{\delta}\|_{1,h}\\&\le\big(\|V\|_\infty+C_0+2\max\{C_0-\tfrac{1}{4},0\}^2\big)\|\boldsymbol{\delta}\|_2^2+\tfrac{1}{8}\langle\boldsymbol{\delta},-\Delta_h\boldsymbol{\delta}\rangle_h.\end{aligned} \tag{4.13}$$

For the last step, let $\boldsymbol{\delta}\neq 0$ and $a:=\|\boldsymbol{\delta}\|_{1,h}/\|\boldsymbol{\delta}\|_2\ge 1$. Then $\langle\boldsymbol{\delta},-\Delta_h\boldsymbol{\delta}\rangle_h=(a^2-1)\|\boldsymbol{\delta}\|_2^2$ by the definition (2.7) of $\|\cdot\|_{1,h}$, and $C_0\|\boldsymbol{\delta}\|_2\|\boldsymbol{\delta}\|_{1,h}-\frac{1}{8}\langle\boldsymbol{\delta},-\Delta_h\boldsymbol{\delta}\rangle_h=\big(C_0a-\frac{a^2-1}{8}\big)\|\boldsymbol{\delta}\|_2^2$. The concave quadratic $a\mapsto C_0a-\frac{a^2-1}{8}$ attains its maximum over $a\ge 1$ at $a=\max\{4C_0,1\}$, where it equals $C_0+2\max\{C_0-\frac{1}{4},0\}^2$. By (3.12), L_0 is the smaller of the two numbers $\|V\|_\infty+3\beta U_0$ and $\|V\|_\infty+C_0+2\max\{C_0-\frac{1}{4},0\}^2$, thus (C3), i.e., $\gamma L_0\le\frac{1}{8}$, means that $\gamma(\|V\|_\infty+3\beta U_0)\le\frac{1}{8}$ if $3\beta U_0\le C_0+2\max\{C_0-\frac{1}{4},0\}^2$, and that $\gamma\big(\|V\|_\infty+C_0+2\max\{C_0-\frac{1}{4},0\}^2\big)\le\frac{1}{8}$ otherwise. In the first case (4.12) multiplied by γ gives $\gamma\langle\mathbf{n}'(\mathbf{u}+\sigma\boldsymbol{\delta})\boldsymbol{\delta},\boldsymbol{\delta}\rangle_h\le\frac{1}{8}\|\boldsymbol{\delta}\|_2^2$, and in the second case (4.13) multiplied by γ gives $\gamma\langle\mathbf{n}'(\mathbf{u}+\sigma\boldsymbol{\delta})\boldsymbol{\delta},\boldsymbol{\delta}\rangle_h\le\frac{1}{8}\|\boldsymbol{\delta}\|_2^2+\frac{1}{8}\gamma\langle\boldsymbol{\delta},-\Delta_h\boldsymbol{\delta}\rangle_h$. In both cases $\gamma\langle\mathbf{n}'(\mathbf{u}+\sigma\boldsymbol{\delta})\boldsymbol{\delta},\boldsymbol{\delta}\rangle_h\le\frac{1}{8}\|\boldsymbol{\delta}\|_{\mathcal{G}}^2$ by (3.7). Integrating against $(1-\sigma)$ in (4.2) halves this bound, which is the first inequality. For the second bound, (3.13) and the componentwise factorization $(\mathbf{u}^+)^3-\mathbf{u}^3=\big((\mathbf{u}^+)^2+\mathbf{u}\mathbf{u}^++\mathbf{u}^2\big)\boldsymbol{\delta}$ give $\mathbf{n}(\mathbf{u}^+)-\mathbf{n}(\mathbf{u})=\mathbb{V}\boldsymbol{\delta}+\beta\big((\mathbf{u}^+)^2+\mathbf{u}\mathbf{u}^++\mathbf{u}^2\big)\boldsymbol{\delta}$, and since componentwise products commute,

$$\langle \mathbf{n}(\mathbf{u}^+)-\mathbf{n}(\mathbf{u}),\mathbf{u}^+\rangle_h=\langle\boldsymbol{\delta},\mathbb{V}\mathbf{u}^+\rangle_h+\beta\langle\boldsymbol{\delta},(\mathbf{u}^+)^3\rangle_h+\beta\langle\boldsymbol{\delta},\mathbf{u}(\mathbf{u}^+)^2\rangle_h+\beta\langle\boldsymbol{\delta},\mathbf{u}^2\mathbf{u}^+\rangle_h.$$

By the Cauchy–Schwarz inequality and $\|\mathbb{V}\mathbf{u}^+\|_2\le\|V\|_\infty\|\mathbf{u}^+\|_2=\|V\|_\infty$,

$$\big|\langle \mathbf{n}(\mathbf{u}^+)-\mathbf{n}(\mathbf{u}),\mathbf{u}^+\rangle_h\big|\le\|\boldsymbol{\delta}\|_2\left(\|V\|_\infty+\beta\left\|(\mathbf{u}^+)^3\right\|_2+\beta\left\|\mathbf{u}(\mathbf{u}^+)^2\right\|_2+\beta\left\|\mathbf{u}^2\mathbf{u}^+\right\|_2\right).$$

Each of the three mixed cubes is bounded in two ways. By Hölder's inequality with exponents 3 and $\frac{3}{2}$ and Lemma 3.2(i) at $\mathbf{u}$ and $\mathbf{u}^+$,

$$\left\|\mathbf{u}(\mathbf{u}^+)^2\right\|_2^2=\sum_i w_iu_i^2(u_i^+)^4\le\Big(\sum_i w_iu_i^6\Big)^{1/3}\Big(\sum_i w_i(u_i^+)^6\Big)^{2/3}=\|\mathbf{u}\|_6^2\left\|\mathbf{u}^+\right\|_6^4\le P_0^3,$$

and in the same way $\left\|\mathbf{u}^2\mathbf{u}^+\right\|_2^2\le\|\mathbf{u}\|_6^4\|\mathbf{u}^+\|_6^2\le P_0^3$ and $\left\|(\mathbf{u}^+)^3\right\|_2^2=\|\mathbf{u}^+\|_6^6\le P_0^3$. By Lemma 3.2(ii) at $\mathbf{u}$ and $\mathbf{u}^+$ and $\|\mathbf{u}\|_2=\|\mathbf{u}^+\|_2=1$,

$$\left\|\mathbf{u}(\mathbf{u}^+)^2\right\|_2\le\|\mathbf{u}\|_\infty\left\|\mathbf{u}^+\right\|_\infty\left\|\mathbf{u}^+\right\|_2\le U_0,\qquad\left\|\mathbf{u}^2\mathbf{u}^+\right\|_2\le\|\mathbf{u}\|_\infty\left\|\mathbf{u}^+\right\|_\infty\|\mathbf{u}\|_2\le U_0,$$
$$\left\|(\mathbf{u}^+)^3\right\|_2\le\left\|\mathbf{u}^+\right\|_\infty^2\left\|\mathbf{u}^+\right\|_2\le U_0.$$

Thus each mixed cube is at most $\min\{P_0^{3/2},U_0\}$, the bracket is at most $\|V\|_\infty+3\beta\min\{P_0^{3/2},U_0\}\le 3B_0$, and the second bound follows from $B_0\le D$ and $\|\boldsymbol{\delta}\|_2\le\|\boldsymbol{\delta}\|_{\mathcal{G}}$. ☐

4.3. **Positivity and the radial interval.**

Lemma 4.4 (Positivity and the radial interval)**.** *Let Assumption 2.1 and (C2), (C4) in* (3.15) *hold, and assume* $\mathbf{z}\neq 0$ *and*

$$\mathbf{u}>0,\qquad \mathbf{E}_h(\mathbf{u})\le E_0,\qquad 1-3\gamma D\le\|\mathbf{z}\|_2\le 1. \tag{4.14}$$

Then $\mathbf{z}^+>0$ *componentwise, and*

$$1-3\gamma D\le\langle\mathbf{u},\mathbf{z}^+\rangle_h\le\left\|\mathbf{z}^+\right\|_2\le 1,\qquad\|\boldsymbol{\delta}\|_2^2\le 2\big(1-\langle\mathbf{u},\mathbf{z}^+\rangle_h\big)\le 6\gamma D,\qquad |R|\le 3\gamma D. \tag{4.15}$$

Proof. We first prove the positivity of $\mathbf{z}^+$. By the first identity of (4.4), $\mathbb{G}_X\mathbf{z}^+=\mathbf{u}-\gamma\mathbf{n}(\mathbf{u})-\gamma(1-\|\mathbf{z}\|_2)(-\Delta_h\mathbf{u})$, and we bound its components from below. By (3.13), $n_i(\mathbf{u})=(V_i+\beta u_i^2)u_i$, and $V_i\le\|V\|_\infty$, $u_i^2\le\|\mathbf{u}\|_\infty^2\le U_0$ by Lemma 3.2(ii) and $u_i>0$ give $n_i(\mathbf{u})\le(\|V\|_\infty+\beta U_0)u_i=B_1u_i$ by (3.11). Since $-\Delta_h=\mathbb{M}^{-1}\mathbb{S}$ with $\mathbb{M}=\operatorname{diag}\{w_1,\dots,w_N\}$, we have $(-\Delta_h\mathbf{u})_i=w_i^{-1}\big(\mathbb{S}_{ii}u_i+\sum_{j\neq i}\mathbb{S}_{ij}u_j\big)\le w_i^{-1}\mathbb{S}_{ii}u_i\le\Lambda_{\mathrm{d}}u_i$, since $\mathbb{S}_{ij}\le 0$ for $j\neq i$ by Lemma 2.2(ii), $\mathbf{u}>0$, and $\Lambda_{\mathrm{d}}=\max_i\mathbb{S}_{ii}/w_i$. With $0\le 1-\|\mathbf{z}\|_2\le 3\gamma D$ from (4.14), we get

$$(\mathbb{G}_X\mathbf{z}^+)_i=u_i-\gamma n_i(\mathbf{u})-\gamma(1-\|\mathbf{z}\|_2)(-\Delta_h\mathbf{u})_i\;\ge\;u_i-\gamma B_1u_i-3\gamma^2D\Lambda_{\mathrm{d}}u_i=\big(1-\gamma B_1-3\gamma^2D\Lambda_{\mathrm{d}}\big)u_i\;\ge\;0$$

for every i, by (C4), i.e., $\mathbb{G}_X\mathbf{z}^+\ge 0$. Since $\mathbb{G}_X^{-1}=(\mathbb{M}+\gamma\mathbb{S})^{-1}\mathbb{M}$ has strictly positive entries by Lemma 2.2(ii) and $w_i>0$, and a matrix with strictly positive entries maps every nonnegative nonzero vector to a positive vector, $\mathbf{z}^+=\mathbb{G}_X^{-1}(\mathbb{G}_X\mathbf{z}^+)>0$ whenever $\mathbf{z}^+\neq 0$. Suppose that $\mathbf{z}^+=0$. Taking the $\langle\cdot,\cdot\rangle_h$-product of the first identity of (4.4) with $\mathbf{u}$ and using $\|\mathbf{u}\|_2=1$, $\langle\mathbf{n}(\mathbf{u}),\mathbf{u}\rangle_h\le B_0\le D$ by Lemma 3.2(iii), $1-\|\mathbf{z}\|_2\le 3\gamma D$ and $\langle\mathbf{u},-\Delta_h\mathbf{u}\rangle_h\le 2E_0\le 2D$ by Lemma 3.2(i), we get

$$0=\langle\mathbf{u},\mathbb{G}_X\mathbf{z}^+\rangle_h=1-\gamma\langle\mathbf{n}(\mathbf{u}),\mathbf{u}\rangle_h-\gamma(1-\|\mathbf{z}\|_2)\langle\mathbf{u},-\Delta_h\mathbf{u}\rangle_h\ge 1-\gamma D-6\gamma^2D^2>0,$$

where the last step holds since $\gamma D\le\frac{1}{128}$ by (C2), a contradiction. Thus $\mathbf{z}^+\neq 0$ and $\mathbf{z}^+>0$.

Next we prove the upper bound $\|\mathbf{z}^+\|_2 \le 1$. By (1.6), $\mathbf{z}^+ = \mathbf{z}-\mathbf{u}+\mathbf{x}$ with $\mathbf{x} = \mathbb{G}_X^{-1}\big(2\mathbf{u}-\mathbf{z}-\gamma\mathbf{n}(\mathbf{u})\big)$, and $\mathbf{z} = \|\mathbf{z}\|_2\,\mathbf{u}$ gives $\mathbf{z}-\mathbf{u} = (\|\mathbf{z}\|_2-1)\mathbf{u}$ and $2\mathbf{u}-\mathbf{z} = (2-\|\mathbf{z}\|_2)\mathbf{u}$, thus

$$\mathbf{z}^+ = (\|\mathbf{z}\|_2-1)\mathbf{u} + (2-\|\mathbf{z}\|_2)\mathbb{G}_X^{-1}\mathbf{u} - \gamma\mathbb{G}_X^{-1}\mathbf{n}(\mathbf{u}). \tag{4.16}$$

Since $V_i \ge 0$ and $\mathbf{u} > 0$, we have $\mathbf{n}(\mathbf{u}) \ge 0$ by (3.13), and $\mathbb{G}_X^{-1} > 0$ entrywise as shown above, thus $\mathbb{G}_X^{-1}\mathbf{n}(\mathbf{u}) \ge 0$ and $0 < \mathbf{z}^+ \le \big((\|\mathbf{z}\|_2-1)\mathbb{I}+(2-\|\mathbf{z}\|_2)\mathbb{G}_X^{-1}\big)\mathbf{u}$ componentwise. Since the weighted norm $\|\cdot\|_2$ is monotone on nonnegative vectors, $\|\mathbf{z}^+\|_2$ is at most the norm of the right side. In the $\langle\cdot,\cdot\rangle_h$-orthonormal eigenbasis of §3.2, in which $\mathbb{G}_X^{-1}$ has the eigenvalues $1/(1+\gamma\mu_j) \in (0,1]$, the operator in parentheses has the eigenvalues $(\|\mathbf{z}\|_2-1)+(2-\|\mathbf{z}\|_2)/(1+\gamma\mu_j)$, which lie in $[\|\mathbf{z}\|_2-1,1] \subset [-1,1]$ for $0 < \|\mathbf{z}\|_2 \le 1$, since $2-\|\mathbf{z}\|_2 > 0$. Thus, with $\hat{u}_j$ the coefficients of $\mathbf{u}$ in that basis,

$$\|\mathbf{z}^+\|_2^2 \le \sum_j \Big((\|\mathbf{z}\|_2-1) + \frac{2-\|\mathbf{z}\|_2}{1+\gamma\mu_j}\Big)^2 \hat{u}_j^2 \le \sum_j \hat{u}_j^2 = \|\mathbf{u}\|_2^2 = 1.$$

Finally we prove the lower bound. Taking the $\langle\cdot,\cdot\rangle_h$-product of (4.16) with $\mathbf{u}$, and using $\|\mathbf{u}\|_2 = 1$, $\langle\mathbf{u},\mathbb{G}_X^{-1}\mathbf{u}\rangle_h = \|\mathbf{u}\|_X^2$ by (3.4) and the h-self-adjointness of $\mathbb{G}_X^{-1}$ in the last term, we have $\langle\mathbf{u},\mathbf{z}^+\rangle_h = (\|\mathbf{z}\|_2-1)+(2-\|\mathbf{z}\|_2)\,\|\mathbf{u}\|_X^2 - \gamma\langle\mathbb{G}_X^{-1}\mathbf{u},\mathbf{n}(\mathbf{u})\rangle_h$, and since $1-(\|\mathbf{z}\|_2-1) = 2-\|\mathbf{z}\|_2$,

$$1-\langle\mathbf{u},\mathbf{z}^+\rangle_h = (2-\|\mathbf{z}\|_2)\big(1-\|\mathbf{u}\|_X^2\big) + \gamma\langle\mathbb{G}_X^{-1}\mathbf{u},\mathbf{n}(\mathbf{u})\rangle_h.$$

Both terms are nonnegative, the first by (3.6) and the second since $\mathbb{G}_X^{-1}\mathbf{u} > 0$ and $\mathbf{n}(\mathbf{u}) \ge 0$. By (3.6) and Lemma 3.2(i), we have $1-\|\mathbf{u}\|_X^2 \le \gamma\langle\mathbf{u},-\Delta_h\mathbf{u}\rangle_h \le 2\gamma E_0$, and by the Cauchy–Schwarz inequality, (3.5) and Lemma 3.2(iii), $\gamma\langle\mathbb{G}_X^{-1}\mathbf{u},\mathbf{n}(\mathbf{u})\rangle_h \le \gamma\,\big\|\mathbb{G}_X^{-1}\mathbf{u}\big\|_2\,\|\mathbf{n}(\mathbf{u})\|_2 \le \gamma\,\|\mathbf{u}\|_X\,\|\mathbf{n}(\mathbf{u})\|_2 \le \gamma B_0$. Therefore, with $2-\|\mathbf{z}\|_2 = 1+(1-\|\mathbf{z}\|_2)$, $1-\|\mathbf{z}\|_2 \le 3\gamma D$, $2E_0+B_0 \le 2D$ and $E_0 \le D$,

$$0 \le 1-\langle\mathbf{u},\mathbf{z}^+\rangle_h \le \gamma(2E_0+B_0) + 2\gamma E_0\,(1-\|\mathbf{z}\|_2) \le 2\gamma D + 6\gamma^2D^2 \le 3\gamma D,$$

where the last step holds since $6\gamma D \le \frac{6}{128} < 1$ by (C2). This proves the first chain in (4.15), since $\langle\mathbf{u},\mathbf{z}^+\rangle_h \le \|\mathbf{u}\|_2\,\|\mathbf{z}^+\|_2 = \|\mathbf{z}^+\|_2$ by the Cauchy–Schwarz inequality. Since $\langle\mathbf{u},\mathbf{z}^+\rangle_h \ge 1-3\gamma D > 0$ and $\|\mathbf{z}^+\|_2 \le 1$, we have $\langle\mathbf{u},\mathbf{z}^+\rangle_h/\,\|\mathbf{z}^+\|_2 \ge \langle\mathbf{u},\mathbf{z}^+\rangle_h$, and (4.1) implies $\|\boldsymbol{\delta}\|_2^2 = 2\big(1-\langle\mathbf{u},\mathbf{z}^+\rangle_h/\,\|\mathbf{z}^+\|_2\big) \le 2\big(1-\langle\mathbf{u},\mathbf{z}^+\rangle_h\big) \le 6\gamma D$. The bound on R follows from the definition (3.8), $R = -(1-\|\mathbf{z}\|_2)+\gamma\lambda(\mathbf{u})$, in which $-(1-\|\mathbf{z}\|_2) \in [-3\gamma D,0]$ by (4.14) and $\gamma\lambda(\mathbf{u}) \in [0,2\gamma D]$ by Lemma 3.2(iv), thus $R \in [-3\gamma D, 2\gamma D]$. ☐

4.4. **Conditional descent.**

Lemma 4.5 (Conditional descent)**.** *Let the hypotheses of Lemma 4.4 hold together with (C3), and assume in addition that* $\mathbf{E}_h(\mathbf{u}^+) \le E_0$. *Then*

$$\gamma\big(\mathcal{L}(\mathbf{z})-\mathcal{L}(\mathbf{z}^+)\big) \;\ge\; \tfrac{5}{16}\,\|\boldsymbol{\delta}\|_{\mathcal{G}}^2 + \tfrac{3}{4}\,R^2, \tag{4.17}$$

and the right side vanishes only if $R = 0$ *and* $\mathbf{g}(\mathbf{u}) = 0$, *in which case* $\mathbf{z}^+ = \mathbf{z}$.

Proof. Since $\mathbf{E}_h(\mathbf{u}) \le E_0$ and $\mathbf{E}_h(\mathbf{u}^+) \le E_0$, Lemma 3.2(i) at $\mathbf{u}$ and $\mathbf{u}^+$ and $E_0 \le D$ give

$$\gamma\langle\mathbf{u},-\Delta_h\mathbf{u}\rangle_h \le 2\gamma E_0 \le 2\gamma D, \qquad \gamma\langle\mathbf{u}^+,-\Delta_h\mathbf{u}^+\rangle_h \le 2\gamma D, \tag{4.18}$$

which are used throughout the proof. By (2.3) and $\Delta_h = -\mathbb{M}^{-1}\mathbb{S}$, we have $\langle\mathbf{v},-\Delta_h\mathbf{w}\rangle_h = \mathbf{v}^\top\mathbb{M}\mathbb{M}^{-1}\mathbb{S}\mathbf{w} = \mathbf{v}^\top\mathbb{S}\mathbf{w}$ for all $\mathbf{v},\mathbf{w}$, and $\mathbb{S}$ is symmetric positive semidefinite by Lemma 2.2(ii), thus $(\mathbf{v},\mathbf{w}) \mapsto \langle\mathbf{v},-\Delta_h\mathbf{w}\rangle_h$ is a symmetric positive semidefinite bilinear form, for which the Cauchy–Schwarz inequality

$$|\langle\mathbf{v},-\Delta_h\mathbf{w}\rangle_h| \le \sqrt{\langle\mathbf{v},-\Delta_h\mathbf{v}\rangle_h}\sqrt{\langle\mathbf{w},-\Delta_h\mathbf{w}\rangle_h}$$

holds. With (4.18) and $\gamma\langle\boldsymbol{\delta}, -\Delta_h\boldsymbol{\delta}\rangle_h \le \|\boldsymbol{\delta}\|_{\mathcal{G}}^2$ from (3.7), we have

$$
\begin{aligned}
\gamma\big|\langle-\Delta_h\mathbf{u}, \boldsymbol{\delta}\rangle_h\big| &\le \sqrt{\gamma\langle\mathbf{u}, -\Delta_h\mathbf{u}\rangle_h}\,\sqrt{\gamma\langle\boldsymbol{\delta}, -\Delta_h\boldsymbol{\delta}\rangle_h} \le \sqrt{2\gamma D}\,\|\boldsymbol{\delta}\|_{\mathcal{G}}\,,\\
\gamma\big|\langle-\Delta_h\boldsymbol{\delta}, \mathbf{u}^+\rangle_h\big| &\le \sqrt{\gamma\langle\mathbf{u}^+, -\Delta_h\mathbf{u}^+\rangle_h}\,\sqrt{\gamma\langle\boldsymbol{\delta}, -\Delta_h\boldsymbol{\delta}\rangle_h} \le \sqrt{2\gamma D}\,\|\boldsymbol{\delta}\|_{\mathcal{G}}\,,\\
\gamma\big|\langle-\Delta_h\mathbf{u}, \mathbf{u}^+\rangle_h\big| &\le \sqrt{\gamma\langle\mathbf{u}, -\Delta_h\mathbf{u}\rangle_h}\,\sqrt{\gamma\langle\mathbf{u}^+, -\Delta_h\mathbf{u}^+\rangle_h} \le 2\gamma D.
\end{aligned}
\tag{4.19}
$$

By (4.15) and $\|\boldsymbol{\delta}\|_2 \le \|\boldsymbol{\delta}\|_{\mathcal{G}}$ from (3.7), we have

$$
\tfrac12\|\boldsymbol{\delta}\|_2^2 \le \tfrac12\|\boldsymbol{\delta}\|_2\sqrt{6\gamma D} \le \sqrt{3\gamma D/2}\,\|\boldsymbol{\delta}\|_{\mathcal{G}}\,, \qquad 0 \le 1 - \left\|\mathbf{z}^+\right\|_2 \le 3\gamma D, \qquad \left\|\mathbf{z}^+\right\|_2 \le 1. \tag{4.20}
$$

Taking absolute values in (4.6), with $1+\gamma\langle\mathbf{u}, -\Delta_h\mathbf{u}\rangle_h \ge 1$, and using (4.19) and (4.20), we get

$$
\begin{aligned}
\big|\left\|\mathbf{z}^+\right\|_2 - \|\mathbf{z}\|_2\big| &\le |R| + \left\|\mathbf{z}^+\right\|_2\left(\gamma\big|\langle-\Delta_h\mathbf{u}, \boldsymbol{\delta}\rangle_h\big| + \tfrac12\|\boldsymbol{\delta}\|_2^2\right)\\
&\le |R| + \big(\sqrt{2}+\sqrt{3/2}\big)\sqrt{\gamma D}\,\|\boldsymbol{\delta}\|_{\mathcal{G}} \le |R| + \tfrac83\sqrt{\gamma D}\,\|\boldsymbol{\delta}\|_{\mathcal{G}}\,,
\end{aligned}
\tag{4.21}
$$

since $\sqrt{2}+\sqrt{3/2} < 2.64 < \frac83$.

We first estimate the energy by (4.9). Since $\|\mathbf{z}^+\|_2 \ge 1-3\gamma D$ by (4.15) and $\|\mathbf{z}^+\|_2 \le 1$, the coefficients of the first two terms on the right side of (4.9) satisfy $\frac{\|\mathbf{z}^+\|_2+1}{2} \ge \|\mathbf{z}^+\|_2 - \frac12 \ge \frac12 - 3\gamma D > 0$, and since $\|\boldsymbol{\delta}\|_2^2 \ge 0$ and $\gamma\langle\boldsymbol{\delta}, -\Delta_h\boldsymbol{\delta}\rangle_h \ge 0$ add up to $\|\boldsymbol{\delta}\|_{\mathcal{G}}^2$ by (3.7), the sum of these two terms is at least $(\frac12 - 3\gamma D)\|\boldsymbol{\delta}\|_{\mathcal{G}}^2$. By (4.19) and (4.21), the third term on the right side of (4.9) is at least

$$
-\big|\left\|\mathbf{z}^+\right\|_2 - \|\mathbf{z}\|_2\big|\,\gamma\big|\langle-\Delta_h\mathbf{u}, \boldsymbol{\delta}\rangle_h\big| \ge -\big|\left\|\mathbf{z}^+\right\|_2 - \|\mathbf{z}\|_2\big|\sqrt{2\gamma D}\,\|\boldsymbol{\delta}\|_{\mathcal{G}} \ge -\sqrt{2\gamma D}\,|R|\,\|\boldsymbol{\delta}\|_{\mathcal{G}} - \tfrac{8\sqrt2}{3}\gamma D\,\|\boldsymbol{\delta}\|_{\mathcal{G}}^2\,,
$$

and Young's inequality $xy \le \frac15 x^2 + \frac54 y^2$ with $x = |R|$ and $y = \sqrt{2\gamma D}\,\|\boldsymbol{\delta}\|_{\mathcal{G}}$ gives $\sqrt{2\gamma D}\,|R|\,\|\boldsymbol{\delta}\|_{\mathcal{G}} \le \frac15 R^2 + \frac52\gamma D\,\|\boldsymbol{\delta}\|_{\mathcal{G}}^2$. With Lemma 4.3 for the remainder term, $-\gamma\mathcal{T}(\mathbf{u},\boldsymbol{\delta}) \ge -\frac1{16}\|\boldsymbol{\delta}\|_{\mathcal{G}}^2$, we get

$$
\gamma\big(\mathbf{E}_h(\mathbf{u}) - \mathbf{E}_h(\mathbf{u}^+)\big) \ge \left(\tfrac12 - \tfrac1{16} - \big(3 + \tfrac{8\sqrt2}{3} + \tfrac52\big)\gamma D\right)\|\boldsymbol{\delta}\|_{\mathcal{G}}^2 - \tfrac15 R^2 \ge \left(\tfrac7{16} - 10\gamma D\right)\|\boldsymbol{\delta}\|_{\mathcal{G}}^2 - \tfrac15 R^2, \tag{4.22}
$$

since $3 + \frac{8\sqrt2}{3} + \frac52 < 9.28 < 10$.

Next we bound the four terms on the right side of (4.10) in absolute value. The first term is bounded by $\sqrt{3\gamma D/2}\,\|\boldsymbol{\delta}\|_{\mathcal{G}}$ via (4.20), the second term by $2\gamma D\big|\|\mathbf{z}^+\|_2 - \|\mathbf{z}\|_2\big| \le 2\gamma D|R| + \frac{16}{3}(\gamma D)^{3/2}\|\boldsymbol{\delta}\|_{\mathcal{G}}$ via (4.19) and (4.21), the third term by $3\gamma D\cdot\sqrt{2\gamma D}\,\|\boldsymbol{\delta}\|_{\mathcal{G}}$ via (4.20) and (4.19), and the fourth term by $3\gamma D\,\|\boldsymbol{\delta}\|_{\mathcal{G}}$ via Lemma 4.3. Adding these bounds, we have

$$
|R^+| \le 2\gamma D|R| + \left(\sqrt{\tfrac32} + \big(\tfrac{16}{3} + 3\sqrt2\big)\gamma D + 3\sqrt{\gamma D}\right)\sqrt{\gamma D}\,\|\boldsymbol{\delta}\|_{\mathcal{G}} \le 2\gamma D|R| + \tfrac85\sqrt{\gamma D}\,\|\boldsymbol{\delta}\|_{\mathcal{G}}\,,
$$

since the bracket is increasing in γD and bounded by $1.2248 + 0.0749 + 0.2652 < 1.57 < \frac85$ at $\gamma D = \frac1{128}$, the largest value allowed by (C2). Squaring with $(x+y)^2 \le 2x^2 + 2y^2$, we obtain

$$
(R^+)^2 \le 2(2\gamma D)^2 R^2 + 2\big(\tfrac85\big)^2\gamma D\,\|\boldsymbol{\delta}\|_{\mathcal{G}}^2 = 8\gamma^2 D^2 R^2 + \tfrac{128}{25}\gamma D\,\|\boldsymbol{\delta}\|_{\mathcal{G}}^2\,. \tag{4.23}
$$

By the definition (3.8) of $\mathcal{L}$, we have $\gamma(\mathcal{L}(\mathbf{z}) - \mathcal{L}(\mathbf{z}^+)) = \gamma(\mathbf{E}_h(\mathbf{u}) - \mathbf{E}_h(\mathbf{u}^+)) + R^2 - (R^+)^2$, thus adding (4.22) and (4.23) we get

$$
\gamma\big(\mathcal{L}(\mathbf{z}) - \mathcal{L}(\mathbf{z}^+)\big) \ge \left(\tfrac7{16} - \big(10 + \tfrac{128}{25}\big)\gamma D\right)\|\boldsymbol{\delta}\|_{\mathcal{G}}^2 + \big(\tfrac45 - 8\gamma^2 D^2\big)R^2 \ge \tfrac5{16}\|\boldsymbol{\delta}\|_{\mathcal{G}}^2 + \tfrac34 R^2,
$$

since $(10 + \frac{128}{25})\gamma D \le 16\gamma D \le \frac18$ and $8\gamma^2 D^2 \le \frac{8}{128^2} < \frac1{20}$ by (C2), and $\frac7{16} - \frac18 = \frac5{16}$, $\frac45 - \frac1{20} = \frac34$.

It remains to prove the equality case. If the right side of (4.17) vanishes, then $\|\boldsymbol{\delta}\|_{\mathcal{G}} = 0$ and $R = 0$. Since $\|\cdot\|_{\mathcal{G}}$ is a norm by (3.7), $\boldsymbol{\delta} = 0$, i.e., $\mathbf{u}^+ = \mathbf{u}$, and (4.8) becomes $\mathbf{z}^+ - \mathbf{z} = (\|\mathbf{z}^+\|_2 - \|\mathbf{z}\|_2)\mathbf{u}$. With $R = 0$, the second identity of (4.4) then becomes $(\|\mathbf{z}^+\|_2 - \|\mathbf{z}\|_2)\,\mathbb{G}_X\mathbf{u} = -\gamma\mathbf{g}(\mathbf{u})$. Taking its $\langle\cdot,\cdot\rangle_h$-product with $\mathbf{u}$ and using $\langle \mathbf{g}(\mathbf{u}),\mathbf{u}\rangle_h = 0$, we get $(\|\mathbf{z}^+\|_2 - \|\mathbf{z}\|_2)\langle \mathbf{u}, \mathbb{G}_X\mathbf{u}\rangle_h = 0$, and since $\langle \mathbf{u}, \mathbb{G}_X\mathbf{u}\rangle_h = \|\mathbf{u}\|_{\mathcal{G}}^2 \ge \|\mathbf{u}\|_2^2 = 1$ by (3.7), we have $\|\mathbf{z}^+\|_2 = \|\mathbf{z}\|_2$, thus $\mathbf{z}^+ = \mathbf{z}$ and $\mathbf{g}(\mathbf{u}) = 0$. □

Lemma 4.6 (The projected gradient and the increment)**.** *Under the hypotheses of Lemma 4.4,*

$$\gamma^2\,\|\mathbf{g}(\mathbf{u})\|_X^2 \le \tfrac{5}{4}\,\|\boldsymbol{\delta}\|_{\mathcal{G}}^2 + 10\gamma D\,R^2, \qquad \left\|\mathbf{z}^+ - \mathbf{z}\right\|_{\mathcal{G}}^2 \le \|\boldsymbol{\delta}\|_{\mathcal{G}}^2 + R^2. \tag{4.24}$$

Proof. Let

$$\boldsymbol{\delta}_\perp := \boldsymbol{\delta} - \frac{\langle \mathbf{u}, \mathbb{G}_X\boldsymbol{\delta}\rangle_h}{\langle \mathbf{u}, \mathbb{G}_X\mathbf{u}\rangle_h}\,\mathbf{u}, \qquad \text{so that} \qquad \langle \mathbf{u}, \mathbb{G}_X\boldsymbol{\delta}_\perp\rangle_h = 0, \qquad \|\boldsymbol{\delta}\|_{\mathcal{G}}^2 = \|\boldsymbol{\delta}_\perp\|_{\mathcal{G}}^2 + \frac{\langle \mathbf{u}, \mathbb{G}_X\boldsymbol{\delta}\rangle_h^2}{\langle \mathbf{u}, \mathbb{G}_X\mathbf{u}\rangle_h} \ge \|\boldsymbol{\delta}_\perp\|_{\mathcal{G}}^2\,,$$

i.e., $\boldsymbol{\delta}_\perp$ is the projection of $\boldsymbol{\delta}$ onto the orthogonal complement of $\mathbf{u}$ with respect to the inner product $\langle\cdot, \mathbb{G}_X\cdot\rangle_h$ of the norm $\|\cdot\|_{\mathcal{G}}$. By (4.1) and the h-self-adjointness of $-\Delta_h$, we have $\langle \mathbf{u}, \mathbb{G}_X\boldsymbol{\delta}\rangle_h = \langle \mathbf{u}, \boldsymbol{\delta}\rangle_h + \gamma\langle \mathbf{u}, -\Delta_h\boldsymbol{\delta}\rangle_h = -\frac{1}{2}\,\|\boldsymbol{\delta}\|_2^2 + \gamma\langle -\Delta_h\mathbf{u}, \boldsymbol{\delta}\rangle_h$ and $\langle \mathbf{u}, \mathbb{G}_X\mathbf{u}\rangle_h = 1 + \gamma\langle \mathbf{u}, -\Delta_h\mathbf{u}\rangle_h$, thus (4.6) reads $\langle \mathbf{u}, \mathbb{G}_X\mathbf{u}\rangle_h\big(\|\mathbf{z}^+\|_2 - \|\mathbf{z}\|_2\big) = -R - \|\mathbf{z}^+\|_2\,\langle \mathbf{u}, \mathbb{G}_X\boldsymbol{\delta}\rangle_h$. Substituting this and $\boldsymbol{\delta} = \boldsymbol{\delta}_\perp + \frac{\langle \mathbf{u}, \mathbb{G}_X\boldsymbol{\delta}\rangle_h}{\langle \mathbf{u}, \mathbb{G}_X\mathbf{u}\rangle_h}\mathbf{u}$ into (4.8), the terms with $\langle \mathbf{u}, \mathbb{G}_X\boldsymbol{\delta}\rangle_h$ cancel and we have

$$\mathbf{z}^+ - \mathbf{z} = \left\|\mathbf{z}^+\right\|_2\boldsymbol{\delta} + (\left\|\mathbf{z}^+\right\|_2 - \|\mathbf{z}\|_2)\mathbf{u} = \left\|\mathbf{z}^+\right\|_2\,\boldsymbol{\delta}_\perp - \frac{R}{\langle \mathbf{u}, \mathbb{G}_X\mathbf{u}\rangle_h}\,\mathbf{u}, \tag{4.25}$$

and substituting (4.25) into the second identity of (4.4), we have

$$\gamma\mathbf{g}(\mathbf{u}) = -\mathbb{G}_X(\mathbf{z}^+ - \mathbf{z}) - R\mathbf{u} = -\left\|\mathbf{z}^+\right\|_2\,\mathbb{G}_X\boldsymbol{\delta}_\perp - R\Big(\mathbf{u} - \frac{\mathbb{G}_X\mathbf{u}}{\langle \mathbf{u}, \mathbb{G}_X\mathbf{u}\rangle_h}\Big). \tag{4.26}$$

The two terms on the right side of (4.25) are orthogonal with respect to $\langle\cdot, \mathbb{G}_X\cdot\rangle_h$, since $\langle \mathbf{u}, \mathbb{G}_X\boldsymbol{\delta}_\perp\rangle_h = 0$, and $\|\mathbf{u}\|_{\mathcal{G}}^2 = \langle \mathbf{u}, \mathbb{G}_X\mathbf{u}\rangle_h$ by (3.7), thus

$$\left\|\mathbf{z}^+ - \mathbf{z}\right\|_{\mathcal{G}}^2 = \left\|\mathbf{z}^+\right\|_2^2\|\boldsymbol{\delta}_\perp\|_{\mathcal{G}}^2 + \frac{R^2}{\langle \mathbf{u}, \mathbb{G}_X\mathbf{u}\rangle_h^2}\,\|\mathbf{u}\|_{\mathcal{G}}^2 = \left\|\mathbf{z}^+\right\|_2^2\|\boldsymbol{\delta}_\perp\|_{\mathcal{G}}^2 + \frac{R^2}{\langle \mathbf{u}, \mathbb{G}_X\mathbf{u}\rangle_h} \le \|\boldsymbol{\delta}\|_{\mathcal{G}}^2 + R^2,$$

by $\|\mathbf{z}^+\|_2 \le 1$ from (4.15), $\|\boldsymbol{\delta}_\perp\|_{\mathcal{G}} \le \|\boldsymbol{\delta}\|_{\mathcal{G}}$ and $\langle \mathbf{u}, \mathbb{G}_X\mathbf{u}\rangle_h = 1 + \gamma\langle \mathbf{u}, -\Delta_h\mathbf{u}\rangle_h \ge 1$, which is the second bound in (4.24). For the first bound, we take the $\|\cdot\|_X$-norm of (4.26). We have $\|\mathbb{G}_X\boldsymbol{\delta}_\perp\|_X = \|\boldsymbol{\delta}_\perp\|_{\mathcal{G}} \le \|\boldsymbol{\delta}\|_{\mathcal{G}}$ by (3.7). By (3.4) and (3.7), $\langle \mathbf{u}, \mathbb{G}_X\mathbf{u}\rangle_X = \langle \mathbf{u}, \mathbb{G}_X^{-1}\mathbb{G}_X\mathbf{u}\rangle_h = \|\mathbf{u}\|_2^2 = 1$ and $\|\mathbb{G}_X\mathbf{u}\|_X^2 = \|\mathbf{u}\|_{\mathcal{G}}^2 = \langle \mathbf{u}, \mathbb{G}_X\mathbf{u}\rangle_h$, thus, expanding the square and using $\|\mathbf{u}\|_X \le \|\mathbf{u}\|_2 = 1$ from (3.5), Lemma 3.2(i) and $E_0 \le D$,

$$\begin{aligned}\left\|\mathbf{u} - \frac{\mathbb{G}_X\mathbf{u}}{\langle \mathbf{u}, \mathbb{G}_X\mathbf{u}\rangle_h}\right\|_X^2 &= \|\mathbf{u}\|_X^2 - \frac{2\langle \mathbf{u}, \mathbb{G}_X\mathbf{u}\rangle_X}{\langle \mathbf{u}, \mathbb{G}_X\mathbf{u}\rangle_h} + \frac{\|\mathbb{G}_X\mathbf{u}\|_X^2}{\langle \mathbf{u}, \mathbb{G}_X\mathbf{u}\rangle_h^2} = \|\mathbf{u}\|_X^2 - \frac{1}{\langle \mathbf{u}, \mathbb{G}_X\mathbf{u}\rangle_h}\\ &\le 1 - \frac{1}{1 + \gamma\langle \mathbf{u}, -\Delta_h\mathbf{u}\rangle_h} = \frac{\gamma\langle \mathbf{u}, -\Delta_h\mathbf{u}\rangle_h}{1 + \gamma\langle \mathbf{u}, -\Delta_h\mathbf{u}\rangle_h} \le \gamma\langle \mathbf{u}, -\Delta_h\mathbf{u}\rangle_h \le 2\gamma D.\end{aligned}$$

By the triangle inequality in (4.26), we get $\gamma\,\|\mathbf{g}(\mathbf{u})\|_X \le \|\mathbf{z}^+\|_2\,\|\boldsymbol{\delta}\|_{\mathcal{G}} + \sqrt{2\gamma D}\,|R| \le \|\boldsymbol{\delta}\|_{\mathcal{G}} + \sqrt{2\gamma D}\,|R|$, and $(x+y)^2 \le \frac{5}{4}x^2 + 5y^2$ implies the first bound. □

4.5. **The energy barrier and the one-step estimates.** Lemma 4.5 assumes that the new direction has energy at most E_0. In this subsection we remove this assumption by a continuation argument, in which the step size and the radial deviation of the input are scaled by the same factor.

Lemma 4.7 (Energy barrier)**.** *Let Assumption 2.1 and (C2)–(C4) in* (3.15) *hold, and assume* $\mathbf{z} \ne 0$, $\mathbf{u} > 0$, $1 - 3\gamma D \le \|\mathbf{z}\|_2 \le 1$ *and* $\mathcal{L}(\mathbf{z}) \le E_0$*. Then* $\mathbf{E}_h(\mathbf{u}^+) \le E_0$*.*

Proof. For $0 \le \theta \le 1$, set

$$\mathbf{z}_\theta := \big(1 - \theta(1 - \|\mathbf{z}\|_2)\big)\mathbf{u}. \tag{4.27}$$

Since $0 \le 1 - \|\mathbf{z}\|_2 \le 3\gamma D < 1$ by the hypothesis $1 - 3\gamma D \le \|\mathbf{z}\|_2 \le 1$ and (C2), the factor $1 - \theta(1 - \|\mathbf{z}\|_2)$ in (4.27) lies in $[1 - 3\theta\gamma D, 1] \subset (0, 1]$, and $\|\mathbf{u}\|_2 = 1$ gives

$$\|\mathbf{z}_\theta\|_2 = 1 - \theta(1 - \|\mathbf{z}\|_2) \in [1 - 3\theta\gamma D, 1], \qquad 1 - \|\mathbf{z}_\theta\|_2 = \theta(1 - \|\mathbf{z}\|_2), \tag{4.28}$$

thus the normalized vector of $\mathbf{z}_\theta$ is $\mathbf{u}$. Let $\mathbf{z}_\theta^+$ denote the result of one step (1.6) of size $\theta\gamma$ from $\mathbf{z}_\theta$, i.e., by the first identity of (4.4) with $\theta\gamma$ in place of γ and (4.28),

$$\mathbf{z}_\theta^+ = (\mathbb{I} - \theta\gamma\Delta_h)^{-1}\big(\mathbf{u} - \theta\gamma\mathbf{n}(\mathbf{u}) - \theta^2\gamma(1 - \|\mathbf{z}\|_2)(-\Delta_h\mathbf{u})\big), \qquad \mathbf{z}_0^+ = \mathbf{u}, \qquad \mathbf{z}_1^+ = \mathbf{z}^+. \tag{4.29}$$

Quantities belonging to the step size $\theta\gamma$ have the subscript θ: $R_\theta(\cdot)$ and $\mathcal{L}_\theta(\cdot)$ are the residual and the Lyapunov function of (3.8) with γ replaced by $\theta\gamma$, and $\mathbf{u}_\theta^+ := \mathbf{z}_\theta^+ / \left\|\mathbf{z}_\theta^+\right\|_2$ if $\mathbf{z}_\theta^+ \neq 0$. Since $\mathbf{z}_1^+ = \mathbf{z}^+$, it suffices to prove that $\mathbf{E}_h(\mathbf{u}_\theta^+) \le E_0$ for every $0 \le \theta \le 1$.

We first show that the continuation is admissible. For $0 < \theta \le 1$ we have $\theta\gamma \le \gamma$, and the left sides of (C2)–(C4) in (3.15) are increasing functions of γ, since the constants D, L_0, B_1 and Λ_{d} of §3.3 do not depend on γ, thus (C2)–(C4) hold with $\theta\gamma$ in place of γ. With (4.28), $\mathbf{u} > 0$ and $\mathbf{E}_h(\mathbf{u}) \le \mathcal{L}(\mathbf{z}) \le E_0$ by (3.8), $\mathbf{z}_\theta$ satisfies (4.14) with $\theta\gamma$ in place of γ. Then Lemma 4.4, applied to $\mathbf{z}_\theta$ with the step size $\theta\gamma$, implies $\mathbf{z}_\theta^+ > 0$ and $\left\|\mathbf{z}_\theta^+\right\|_2 \ge 1 - 3\theta\gamma D > 0$, and $\mathbf{u}_\theta^+$ is defined for every $\theta \in [0, 1]$, with $\mathbf{u}_0^+ = \mathbf{u}$ by (4.29). Since $(\mathbb{I} - \theta\gamma\Delta_h)^{-1} = (\mathbb{M} + \theta\gamma\mathbb{S})^{-1}\mathbb{M}$ and $\mathbb{M} + \theta\gamma\mathbb{S}$ is positive definite for $\theta \ge 0$, the entries of $\mathbf{z}_\theta^+$ in (4.29) are rational functions of θ whose denominators do not vanish on $[0, 1]$, thus $\theta \mapsto \mathbf{E}_h(\mathbf{u}_\theta^+)$ is continuous on $[0, 1]$ and differentiable at $\theta = 0$. By (3.8) with $\theta\gamma$ in place of γ and (4.28), for $0 < \theta \le 1$ the residual and the Lyapunov function of $\mathbf{z}_\theta$ are

$$\begin{aligned} R_\theta(\mathbf{z}_\theta) &= \|\mathbf{z}_\theta\|_2 - 1 + \theta\gamma\lambda(\mathbf{u}) = -\theta(1 - \|\mathbf{z}\|_2) + \theta\gamma\lambda(\mathbf{u}) = \theta R, \\ \mathcal{L}_\theta(\mathbf{z}_\theta) &= \mathbf{E}_h(\mathbf{u}) + \frac{(\theta R)^2}{\theta\gamma} = \mathbf{E}_h(\mathbf{u}) + \frac{\theta R^2}{\gamma} \le \mathbf{E}_h(\mathbf{u}) + \frac{R^2}{\gamma} = \mathcal{L}(\mathbf{z}) \le E_0, \end{aligned} \tag{4.30}$$

where we used $\mathbf{u} = \mathbf{z}_\theta / \|\mathbf{z}_\theta\|_2$ in the computation of the Lyapunov function.

Next we consider the case $R = 0$ and $\mathbf{g}(\mathbf{u}) = 0$. The second identity of (4.4), applied to the step of size $\theta\gamma$ from $\mathbf{z}_\theta$, whose normalized vector is $\mathbf{u}$ and whose residual is θR by (4.30), implies $(\mathbb{I} - \theta\gamma\Delta_h)(\mathbf{z}_\theta^+ - \mathbf{z}_\theta) = -\big(\theta R\mathbf{u} + \theta\gamma\mathbf{g}(\mathbf{u})\big) = -\theta\big(R\mathbf{u} + \gamma\mathbf{g}(\mathbf{u})\big) = 0$, thus $\mathbf{z}_\theta^+ = \mathbf{z}_\theta = \|\mathbf{z}_\theta\|_2\mathbf{u}$, which implies $\mathbf{u}_\theta^+ = \mathbf{u}$ and $\mathbf{E}_h(\mathbf{u}_\theta^+) = \mathbf{E}_h(\mathbf{u}) \le E_0$ for every θ. In the rest of the proof we assume that $R \neq 0$ or $\mathbf{g}(\mathbf{u}) \neq 0$.

We claim that there is $\theta_1 > 0$ such that $\mathbf{E}_h(\mathbf{u}_\theta^+) < E_0$ for $0 < \theta < \theta_1$. If $\mathbf{E}_h(\mathbf{u}) < E_0$, this follows from the continuity of $\theta \mapsto \mathbf{E}_h(\mathbf{u}_\theta^+)$ at $\theta = 0$, where its value is $\mathbf{E}_h(\mathbf{u}_0^+) = \mathbf{E}_h(\mathbf{u})$. If $\mathbf{E}_h(\mathbf{u}) = E_0$, then $R = 0$ since $\mathcal{L}(\mathbf{z}) = \mathbf{E}_h(\mathbf{u}) + R^2/\gamma \le E_0$ by (3.8), and therefore $\mathbf{g}(\mathbf{u}) \neq 0$. Multiplying (4.29) by $\mathbb{I} - \theta\gamma\Delta_h$ gives $(\mathbb{I} - \theta\gamma\Delta_h)\mathbf{z}_\theta^+ = \mathbf{u} - \theta\gamma\mathbf{n}(\mathbf{u}) - \theta^2\gamma(1 - \|\mathbf{z}\|_2)(-\Delta_h\mathbf{u})$, and differentiating both sides with respect to θ by the product rule gives

$$-\gamma\Delta_h\mathbf{z}_\theta^+ + (\mathbb{I} - \theta\gamma\Delta_h)\frac{d}{d\theta}\mathbf{z}_\theta^+ = -\gamma\mathbf{n}(\mathbf{u}) - 2\theta\gamma(1 - \|\mathbf{z}\|_2)(-\Delta_h\mathbf{u}).$$

At $\theta = 0$, where $\mathbf{z}_0^+ = \mathbf{u}$, this reads $-\gamma\Delta_h\mathbf{u} + \frac{d}{d\theta}\mathbf{z}_\theta^+\big|_{\theta=0} = -\gamma\mathbf{n}(\mathbf{u})$, thus, by (3.13),

$$\frac{d}{d\theta}\mathbf{z}_\theta^+\bigg|_{\theta=0} = \gamma\Delta_h\mathbf{u} - \gamma\mathbf{n}(\mathbf{u}) = -\gamma\big(-\Delta_h\mathbf{u} + \mathbf{n}(\mathbf{u})\big) = -\gamma\nabla\mathbf{E}_h(\mathbf{u}).$$

Differentiating $\left\|\mathbf{z}_\theta^+\right\|_2^2 = \langle \mathbf{z}_\theta^+, \mathbf{z}_\theta^+\rangle_h$ gives $\frac{d}{d\theta}\left\|\mathbf{z}_\theta^+\right\|_2 = \langle \mathbf{z}_\theta^+, \frac{d}{d\theta}\mathbf{z}_\theta^+\rangle_h / \left\|\mathbf{z}_\theta^+\right\|_2$, thus the quotient rule applied to $\mathbf{u}_\theta^+ = \mathbf{z}_\theta^+ / \left\|\mathbf{z}_\theta^+\right\|_2$ gives

$$\frac{d}{d\theta}\mathbf{u}_\theta^+ = \frac{1}{\left\|\mathbf{z}_\theta^+\right\|_2}\frac{d}{d\theta}\mathbf{z}_\theta^+ - \frac{\langle \mathbf{z}_\theta^+, \frac{d}{d\theta}\mathbf{z}_\theta^+\rangle_h}{\left\|\mathbf{z}_\theta^+\right\|_2^3}\,\mathbf{z}_\theta^+.$$

At $\theta = 0$, where $\mathbf{z}_0^+ = \mathbf{u}$, $\|\mathbf{u}\|_2 = 1$ and $\frac{d}{d\theta}\mathbf{z}_\theta^+\big|_{\theta=0} = -\gamma\nabla\mathbf{E}_h(\mathbf{u})$, this gives, by the definitions (3.3) of $\lambda(\mathbf{u})$ and $\mathbf{g}(\mathbf{u})$,

$$\frac{d}{d\theta}\mathbf{u}_\theta^+\Big|_{\theta=0} = -\gamma\nabla\mathbf{E}_h(\mathbf{u}) + \gamma\langle \mathbf{u}, \nabla\mathbf{E}_h(\mathbf{u})\rangle_h\,\mathbf{u} = -\gamma\big(\nabla\mathbf{E}_h(\mathbf{u}) - \lambda(\mathbf{u})\mathbf{u}\big) = -\gamma\mathbf{g}(\mathbf{u}),$$

and then, by the chain rule,

$$\frac{d}{d\theta}\mathbf{E}_h(\mathbf{u}_\theta^+)\Big|_{\theta=0} = \big\langle \nabla\mathbf{E}_h(\mathbf{u}), -\gamma\mathbf{g}(\mathbf{u})\big\rangle_h = -\gamma\big\langle \lambda(\mathbf{u})\mathbf{u} + \mathbf{g}(\mathbf{u}), \mathbf{g}(\mathbf{u})\big\rangle_h = -\gamma\,\|\mathbf{g}(\mathbf{u})\|_2^2 < 0,$$

where $\nabla\mathbf{E}_h(\mathbf{u}) = \lambda(\mathbf{u})\mathbf{u} + \mathbf{g}(\mathbf{u})$ by (3.3), $\langle \mathbf{g}(\mathbf{u}), \mathbf{u}\rangle_h = 0$ and $\mathbf{g}(\mathbf{u}) \neq 0$ are used in the last two steps. A function with value E_0 and negative derivative at $\theta = 0$ is below E_0 on some interval $(0, \theta_1)$, which proves the claim.

Suppose that $\mathbf{E}_h(\mathbf{u}_\theta^+) > E_0$ for some $\theta \in (0, 1]$, and let θ_* be the infimum of all such θ. By the claim, every such θ is at least θ_1, thus $\theta_* \geq \theta_1 > 0$. By the definition of the infimum, $\mathbf{E}_h(\mathbf{u}_\theta^+) \leq E_0$ for $0 < \theta < \theta_*$, and by continuity $\mathbf{E}_h(\mathbf{u}_{\theta_*}^+) \leq E_0$ as a limit from the left and $\mathbf{E}_h(\mathbf{u}_{\theta_*}^+) \geq E_0$ as a limit of values above E_0, thus $\mathbf{E}_h(\mathbf{u}_{\theta_*}^+) = E_0$. Then Lemma 4.5 applies to the step of size $\theta_*\gamma$ from $\mathbf{z}_{\theta_*}$, whose new normalized vector is $\mathbf{u}_{\theta_*}^+$, since (C2)–(C4) hold with $\theta_*\gamma$ in place of γ, $\mathbf{z}_{\theta_*}$ satisfies (4.14) with $\theta_*\gamma$ in place of γ, as shown at the beginning of the proof, and $\mathbf{E}_h(\mathbf{u}_{\theta_*}^+) = E_0$. By (3.8), (4.17) and (4.30), we get

$$E_0 = \mathbf{E}_h(\mathbf{u}_{\theta_*}^+) \leq \mathcal{L}_{\theta_*}(\mathbf{z}_{\theta_*}^+) \leq \mathcal{L}_{\theta_*}(\mathbf{z}_{\theta_*}) \leq E_0.$$

Therefore equality holds throughout, in particular $\mathcal{L}_{\theta_*}(\mathbf{z}_{\theta_*}) - \mathcal{L}_{\theta_*}(\mathbf{z}_{\theta_*}^+) = 0$, thus the right side of (4.17) vanishes for that step, and the equality case of Lemma 4.5 gives $R_{\theta_*}(\mathbf{z}_{\theta_*}) = \theta_* R = 0$ and $\mathbf{g}(\mathbf{u}) = 0$. Since $\theta_* > 0$, we get $R = 0$ and $\mathbf{g}(\mathbf{u}) = 0$, which contradicts the assumption that $R \neq 0$ or $\mathbf{g}(\mathbf{u}) \neq 0$. Thus $\mathbf{E}_h(\mathbf{u}_\theta^+) \leq E_0$ for every $\theta \in [0, 1]$. □

Proposition 4.8 (One-step estimates)**.** *Let Assumption 2.1 hold, let* $0 < \gamma \leq \gamma_{\max}$*, and assume* $\mathbf{z} \neq 0$*,* $\mathbf{u} > 0$*,* $1 - 3\gamma D \leq \|\mathbf{z}\|_2 \leq 1$ *and* $\mathcal{L}(\mathbf{z}) \leq E_0$*. Then* $\mathbf{z}^+ > 0$ *componentwise,* $1 - 3\gamma D \leq \|\mathbf{z}^+\|_2 \leq 1$*,* $\mathbf{E}_h(\mathbf{u}^+) \leq E_0$*, and*

$$\mathcal{L}(\mathbf{z}) - \mathcal{L}(\mathbf{z}^+) \;\geq\; \tfrac{1}{4}\,\gamma\,\|\mathbf{g}(\mathbf{u})\|_X^2 + \frac{R^2}{2\gamma}, \qquad \mathcal{L}(\mathbf{z}) - \mathcal{L}(\mathbf{z}^+) \;\geq\; \frac{5}{16\gamma}\,\left\|\mathbf{z}^+ - \mathbf{z}\right\|_{\mathcal{G}}^2. \tag{4.31}$$

Proof. By Lemma 3.1, (C1)–(C4) hold. By (3.8), $\mathbf{E}_h(\mathbf{u}) \leq \mathcal{L}(\mathbf{z}) \leq E_0$, thus $\mathbf{z}$ satisfies (4.14), and Lemma 4.4 implies $\mathbf{z}^+ > 0$ and $1 - 3\gamma D \leq \|\mathbf{z}^+\|_2 \leq 1$, Lemma 4.7 implies $\mathbf{E}_h(\mathbf{u}^+) \leq E_0$, and then Lemma 4.5 implies $\gamma(\mathcal{L}(\mathbf{z}) - \mathcal{L}(\mathbf{z}^+)) \geq \frac{5}{16}\|\boldsymbol{\delta}\|_{\mathcal{G}}^2 + \frac{3}{4}R^2$. The first bound in (4.24) multiplied by $\frac{1}{4}$ gives $\frac{5}{16}\|\boldsymbol{\delta}\|_{\mathcal{G}}^2 \geq \frac{1}{4}\gamma^2\|\mathbf{g}(\mathbf{u})\|_X^2 - \frac{5}{2}\gamma D\,R^2$, thus

$$\gamma\big(\mathcal{L}(\mathbf{z}) - \mathcal{L}(\mathbf{z}^+)\big) \geq \tfrac{1}{4}\gamma^2\|\mathbf{g}(\mathbf{u})\|_X^2 + \big(\tfrac{3}{4} - \tfrac{5}{2}\gamma D\big)R^2 \geq \tfrac{1}{4}\gamma^2\|\mathbf{g}(\mathbf{u})\|_X^2 + \tfrac{1}{2}R^2,$$

since $\frac{5}{2}\gamma D \leq \frac{5}{256} < \frac{1}{4}$ by (C2), and dividing by γ proves the first inequality in (4.31). The second follows from $\frac{5}{16}\|\boldsymbol{\delta}\|_{\mathcal{G}}^2 + \frac{3}{4}R^2 \geq \frac{5}{16}(\|\boldsymbol{\delta}\|_{\mathcal{G}}^2 + R^2) \geq \frac{5}{16}\|\mathbf{z}^+ - \mathbf{z}\|_{\mathcal{G}}^2$, by the second bound in (4.24), divided by γ. □

5. Global convergence to the ground state

We now prove Theorem 3.4 by induction, and then prove the local linear rate. In particular, the induction is used to prove the following invariance:

$$\mathbf{u}^k > 0, \qquad 1 - 3\gamma D \le \left\|\mathbf{z}^k\right\|_2 \le 1, \qquad \mathcal{L}^k \le E_0. \tag{5.1}$$

The second clause contains $\mathbf{z}^k \neq 0$, so that $\mathbf{u}^k$, R^k and $\mathcal{L}^k$ are defined, and the third clause contains the energy bound $\mathbf{E}_h(\mathbf{u}^k) \le E_0$. The second and third clauses imply, by $R^k = \left\|\mathbf{z}^k\right\|_2 - 1 + \gamma\lambda(\mathbf{u}^k)$ and Lemma 3.2(iv), the residual bound $|R^k| \le 3\gamma D$.

Proof of Theorem 3.4. We first prove by induction that (5.1) holds at every step $k \ge 0$, together with the drop

$$\mathcal{L}^k - \mathcal{L}^{k+1} \;\ge\; \tfrac{1}{4}\,\gamma\left\|\mathbf{g}(\mathbf{u}^k)\right\|_X^2 + \frac{(R^k)^2}{2\gamma}, \qquad \mathcal{L}^k - \mathcal{L}^{k+1} \;\ge\; \frac{5}{16\gamma}\left\|\mathbf{z}^{k+1} - \mathbf{z}^k\right\|_{\mathcal{G}}^2. \tag{5.2}$$

At $k = 0$, $\mathbf{u}^0 > 0$ and $\left\|\mathbf{z}^0\right\|_2 = 1$ by hypothesis, and $\mathcal{L}^0 \le E_0$ by Lemma 3.1, thus (5.1) holds at step 0. Suppose that (5.1) holds at step k. Then $\mathbf{z}^k$ satisfies the hypotheses of Proposition 4.8, which implies $\mathbf{z}^{k+1} > 0$, thus $\mathbf{u}^{k+1} = \mathbf{z}^{k+1}/\left\|\mathbf{z}^{k+1}\right\|_2 > 0$, the radial interval $1 - 3\gamma D \le \left\|\mathbf{z}^{k+1}\right\|_2 \le 1$, and the drop (5.2) at step k, whose right sides are nonnegative, thus $\mathcal{L}^{k+1} \le \mathcal{L}^k \le E_0$. Hence (5.1) holds at step $k+1$. By induction, (5.1) and (5.2) hold at every step $k \ge 0$. In particular $\mathbf{u}^k > 0$ and $\mathbf{z}^k \neq 0$ for every $k \ge 0$, so the whole sequence is well defined.

Next we prove the summability of the descent terms. Summing (5.2) over $0 \le k \le m$, the left side equals $\mathcal{L}^0 - \mathcal{L}^{m+1} \le \mathcal{L}^0$, since $\mathcal{L}^{m+1} \ge \mathbf{E}_h(\mathbf{u}^{m+1}) \ge 0$, every term of $\mathbf{E}_h$ being nonnegative under Assumption 2.1. Thus for every m we have

$$\tfrac{1}{4}\sum_{k=0}^{m}\gamma\left\|\mathbf{g}^k\right\|_X^2 + \frac{1}{2}\sum_{k=0}^{m}\frac{(R^k)^2}{\gamma} \;\le\; \mathcal{L}^0, \qquad \frac{5}{16\gamma}\sum_{k=0}^{m}\left\|\mathbf{z}^{k+1} - \mathbf{z}^k\right\|_{\mathcal{G}}^2 \;\le\; \mathcal{L}^0,$$

and letting $m \to \infty$, we get

$$\sum_{k\ge 0}\gamma\left\|\mathbf{g}^k\right\|_X^2 \le 4\,\mathcal{L}^0, \qquad \sum_{k\ge 0}\frac{(R^k)^2}{\gamma} \le 2\,\mathcal{L}^0, \qquad \sum_{k\ge 0}\left\|\mathbf{z}^{k+1} - \mathbf{z}^k\right\|_{\mathcal{G}}^2 \le \tfrac{16}{5}\,\gamma\,\mathcal{L}^0.$$

The terms of these convergent series tend to zero and $\gamma > 0$ is fixed, thus $\left\|\mathbf{g}^k\right\|_X \to 0$ and $R^k \to 0$. With $\mu_{\max}$ the largest eigenvalue of $-\Delta_h$, the spectral representation of $\|\cdot\|_X$ in §3.2 gives, for every $\mathbf{v} \in \mathbb{R}^N$,

$$\|\mathbf{v}\|_2^2 = \sum_j \hat{v}_j^2 \le (1 + \gamma\mu_{\max})\sum_j \frac{\hat{v}_j^2}{1 + \gamma\mu_j} = (1 + \gamma\mu_{\max})\,\|\mathbf{v}\|_X^2\,, \tag{5.3}$$

thus also $\left\|\mathbf{g}^k\right\|_2 \to 0$.

Next we consider the accumulation points. The manifold $\mathcal{M}$ is a closed bounded subset of $\mathbb{R}^N$, thus compact, so $(\mathbf{u}^k)$ has accumulation points. Let $\mathbf{u}^{k_j} \to \mathbf{u}^*$. Each component of $\mathbf{g}(\mathbf{u}) = \nabla\mathbf{E}_h(\mathbf{u}) - \lambda(\mathbf{u})\mathbf{u}$ in (3.3) is a polynomial in the components of $\mathbf{u}$, so $\mathbf{g}$ is continuous and $\mathbf{g}(\mathbf{u}^{k_j}) \to \mathbf{g}(\mathbf{u}^*)$, while $\left\|\mathbf{g}(\mathbf{u}^k)\right\|_2 \to 0$ by the previous paragraph. Thus $\mathbf{g}(\mathbf{u}^*) = 0$, so $\mathbf{u}^*$ is a critical point of $\mathbf{E}_h|_{\mathcal{M}}$. By (5.1) we have $\mathbf{u}^k > 0$ for every $k \ge 0$, and passing to the limit along $\mathbf{u}^{k_j} \to \mathbf{u}^*$ we get $\mathbf{u}^* \ge 0$. The strict positivity of $\mathbf{u}^*$ will follow from Remark 2.3.

Finally we identify the limit. Every accumulation point $\mathbf{u}^*$ is thus a nonnegative critical point of $\mathbf{E}_h|_{\mathcal{M}}$, and by Remark 2.3 the only such point is $\mathbf{u}_{\rm GS}$. Thus $\mathbf{u}^* = \mathbf{u}_{\rm GS}$ for every accumulation point. Suppose that $\mathbf{u}^k \not\to \mathbf{u}_{\rm GS}$. Then there are $\epsilon > 0$ and a subsequence $(\mathbf{u}^{k_j})$ with $\left\|\mathbf{u}^{k_j} - \mathbf{u}_{\rm GS}\right\|_2 \ge \epsilon$ for every j, which has an accumulation point by the compactness of $\mathcal{M}$, and this accumulation point is at distance at least ϵ from $\mathbf{u}_{\mathrm{GS}}$, a contradiction. Thus $\mathbf{u}^k \to \mathbf{u}_{\mathrm{GS}}$. Since $\mathbf{E}_h$ is continuous, we get $\mathbf{E}_h(\mathbf{u}^k) \to \mathbf{E}_h(\mathbf{u}_{\mathrm{GS}}) = \min_{\mathcal{M}} \mathbf{E}_h$, the equality again by Remark 2.3. $\square$

Once the iterates are close to the ground state, the descent inequality (5.2) implies a linear rate. The rate is a statement for the fixed mesh, and neither the index from which it holds nor its constant is uniform in h.

Theorem 5.1 (Local linear convergence)**.** *Under the hypotheses of Theorem 3.4 and with $N \geq 2$, write $E_{\mathrm{GS}} := \mathbf{E}_h(\mathbf{u}_{\mathrm{GS}})$, $\lambda_{\mathrm{GS}} := \lambda(\mathbf{u}_{\mathrm{GS}})$ and $m_{\mathrm{GS}} := \min_i(\mathbf{u}_{\mathrm{GS}})_i > 0$, let $\mathbf{n}'$ be as in* (4.2) *and $A_{\mathbf{u}}$ as in* (2.6)*, let*

$$\mathcal{H} := -\Delta_h + \mathbf{n}'(\mathbf{u}_{\mathrm{GS}}) - \lambda_{\mathrm{GS}}\mathbb{I} = A_{\mathbf{u}_{\mathrm{GS}}} - \lambda_{\mathrm{GS}}\mathbb{I} + 2\beta\,\mathrm{diag}(\mathbf{u}_{\mathrm{GS}}^2), \tag{5.4}$$

and let ν be the minimum of $\langle \mathbf{w}, \mathcal{H}\mathbf{w}\rangle_h$ over all $\mathbf{w} \in \mathcal{M}$ with $\langle \mathbf{u}_{\mathrm{GS}}, \mathbf{w}\rangle_h = 0$. Then $\nu \geq 2\beta m_{\mathrm{GS}}^2 > 0$, and with $\mu_{\max}$ the largest eigenvalue of $-\Delta_h$ and

$$\eta := \min\Big\{\frac{\gamma\nu}{4(1+\gamma\mu_{\max})},\ \frac{1}{2}\Big\} \in (0,1), \tag{5.5}$$

there is an index K such that, for every $k \geq K$,

$$\begin{gathered} \mathcal{L}^{k+1} - E_{\mathrm{GS}} \leq (1-\eta)\big(\mathcal{L}^k - E_{\mathrm{GS}}\big), \\ \mathbf{E}_h(\mathbf{u}^k) - E_{\mathrm{GS}} + \frac{(R^k)^2}{\gamma} + \frac{\nu}{4}\left\|\mathbf{u}^k - \mathbf{u}_{\mathrm{GS}}\right\|_2^2 \leq 2\,(1-\eta)^{k-K}\big(\mathcal{L}^K - E_{\mathrm{GS}}\big). \end{gathered} \tag{5.6}$$

Proof. We first prove the lower bound on ν. By Appendix A, $\mathbf{u}_{\mathrm{GS}} > 0$ is an eigenvector of the h-self-adjoint operator $A_{\mathbf{u}_{\mathrm{GS}}}$ belonging to its smallest eigenvalue. Since $\nabla\mathbf{E}_h(\mathbf{u}) = A_{\mathbf{u}}\mathbf{u}$ by §3.1 and $\mathbf{g}(\mathbf{u}_{\mathrm{GS}}) = 0$, (3.3) gives $A_{\mathbf{u}_{\mathrm{GS}}}\mathbf{u}_{\mathrm{GS}} = \nabla\mathbf{E}_h(\mathbf{u}_{\mathrm{GS}}) = \lambda_{\mathrm{GS}}\mathbf{u}_{\mathrm{GS}}$, thus this smallest eigenvalue is λ_{GS}, and since the Rayleigh quotient of an h-self-adjoint operator is at least its smallest eigenvalue, $\langle \mathbf{w}, A_{\mathbf{u}_{\mathrm{GS}}}\mathbf{w}\rangle_h \geq \lambda_{\mathrm{GS}}\left\|\mathbf{w}\right\|_2^2$ for every $\mathbf{w}$. By (5.4) and $(\mathbf{u}_{\mathrm{GS}})_i^2 \geq m_{\mathrm{GS}}^2$ for every i, we get

$$\begin{aligned} \langle \mathbf{w}, \mathcal{H}\mathbf{w}\rangle_h &= \langle \mathbf{w}, A_{\mathbf{u}_{\mathrm{GS}}}\mathbf{w}\rangle_h - \lambda_{\mathrm{GS}}\left\|\mathbf{w}\right\|_2^2 + 2\beta\langle \mathbf{w}, \mathrm{diag}(\mathbf{u}_{\mathrm{GS}}^2)\mathbf{w}\rangle_h \\ &\geq 2\beta\langle \mathbf{w}, \mathrm{diag}(\mathbf{u}_{\mathrm{GS}}^2)\mathbf{w}\rangle_h \geq 2\beta m_{\mathrm{GS}}^2\left\|\mathbf{w}\right\|_2^2, \end{aligned}$$

and taking the minimum over $\mathbf{w} \in \mathcal{M}$ with $\langle \mathbf{u}_{\mathrm{GS}}, \mathbf{w}\rangle_h = 0$ gives $\nu \geq 2\beta m_{\mathrm{GS}}^2$.

Next we introduce a chart around the ground state. Let $T_{\mathrm{GS}} := \{\boldsymbol{\xi} : \langle \mathbf{u}_{\mathrm{GS}}, \boldsymbol{\xi}\rangle_h = 0\}$ be the tangent space of $\mathcal{M}$ at $\mathbf{u}_{\mathrm{GS}}$ and, for $\boldsymbol{\xi} \in T_{\mathrm{GS}}$,

$$\mathbf{v}(\boldsymbol{\xi}) := \frac{\mathbf{u}_{\mathrm{GS}} + \boldsymbol{\xi}}{\sqrt{1 + \left\|\boldsymbol{\xi}\right\|_2^2}} \in \mathcal{M}, \qquad f(\boldsymbol{\xi}) := \mathbf{E}_h(\mathbf{v}(\boldsymbol{\xi})),$$

where $\left\|\mathbf{u}_{\mathrm{GS}} + \boldsymbol{\xi}\right\|_2^2 = 1 + \left\|\boldsymbol{\xi}\right\|_2^2$ by $\left\|\mathbf{u}_{\mathrm{GS}}\right\|_2 = 1$ and $\langle \mathbf{u}_{\mathrm{GS}}, \boldsymbol{\xi}\rangle_h = 0$. Every $\mathbf{u} \in \mathcal{M}$ with $\langle \mathbf{u}, \mathbf{u}_{\mathrm{GS}}\rangle_h > 0$ is of this form: for $\boldsymbol{\xi} = \mathbf{u}/\langle \mathbf{u}, \mathbf{u}_{\mathrm{GS}}\rangle_h - \mathbf{u}_{\mathrm{GS}}$ we have $\langle \mathbf{u}_{\mathrm{GS}}, \boldsymbol{\xi}\rangle_h = 1 - 1 = 0$, $\left\|\boldsymbol{\xi}\right\|_2^2 = \langle \mathbf{u}, \mathbf{u}_{\mathrm{GS}}\rangle_h^{-2} - 2 + 1 = \langle \mathbf{u}, \mathbf{u}_{\mathrm{GS}}\rangle_h^{-2} - 1$ and $\mathbf{v}(\boldsymbol{\xi}) = \big(\mathbf{u}/\langle \mathbf{u}, \mathbf{u}_{\mathrm{GS}}\rangle_h\big)/\sqrt{\langle \mathbf{u}, \mathbf{u}_{\mathrm{GS}}\rangle_h^{-2}} = \mathbf{u}$. By $(1+p)^{-1/2} = 1 - \frac{p}{2} + O(p^2)$ with $p = \left\|\boldsymbol{\xi}\right\|_2^2$,

$$\mathbf{v}(\boldsymbol{\xi}) - \mathbf{u}_{\mathrm{GS}} = \boldsymbol{\xi} - \tfrac{1}{2}\left\|\boldsymbol{\xi}\right\|_2^2\mathbf{u}_{\mathrm{GS}} + O(\left\|\boldsymbol{\xi}\right\|_2^3).$$

The gradient of $\mathbf{E}_h$ is $\nabla\mathbf{E}_h(\mathbf{u}) = -\Delta_h\mathbf{u} + \mathbf{n}(\mathbf{u})$ by (3.13), and the Jacobian of $\mathbf{n}$ is $\mathbf{n}'$ of (4.2), thus the Hessian of $\mathbf{E}_h$ with respect to $\langle\cdot,\cdot\rangle_h$ is $-\Delta_h + \mathbf{n}'(\mathbf{u})$. Since $\mathbf{E}_h$ is a polynomial of degree four, its Taylor expansion at $\mathbf{u}_{\rm GS}$ reads, for every $\mathbf{v}\in\mathbb{R}^N$,

$$\begin{aligned}\mathbf{E}_h(\mathbf{v}) = E_{\rm GS} &+ \langle \nabla\mathbf{E}_h(\mathbf{u}_{\rm GS}), \mathbf{v}-\mathbf{u}_{\rm GS}\rangle_h \\ &+ \tfrac{1}{2}\big\langle \mathbf{v}-\mathbf{u}_{\rm GS}, \big(-\Delta_h + \mathbf{n}'(\mathbf{u}_{\rm GS})\big)(\mathbf{v}-\mathbf{u}_{\rm GS})\big\rangle_h + O(\|\mathbf{v}-\mathbf{u}_{\rm GS}\|_2^3).\end{aligned}$$

We take $\mathbf{v}=\mathbf{v}(\boldsymbol{\xi})$, so that $\mathbf{E}_h(\mathbf{v}) = f(\boldsymbol{\xi})$, $\mathbf{v}-\mathbf{u}_{\rm GS} = \boldsymbol{\xi} - \frac{1}{2}\|\boldsymbol{\xi}\|_2^2\mathbf{u}_{\rm GS} + O(\|\boldsymbol{\xi}\|_2^3)$ and $\|\mathbf{v}-\mathbf{u}_{\rm GS}\|_2 = O(\|\boldsymbol{\xi}\|_2)$. With $\nabla\mathbf{E}_h(\mathbf{u}_{\rm GS}) = \lambda_{\rm GS}\mathbf{u}_{\rm GS}$, $\langle \mathbf{u}_{\rm GS}, \boldsymbol{\xi}\rangle_h = 0$ and $\|\mathbf{u}_{\rm GS}\|_2 = 1$, the linear and the quadratic terms are

$$\begin{aligned}\langle \nabla\mathbf{E}_h(\mathbf{u}_{\rm GS}), \mathbf{v}-\mathbf{u}_{\rm GS}\rangle_h &= \lambda_{\rm GS}\langle \mathbf{u}_{\rm GS}, \mathbf{v}-\mathbf{u}_{\rm GS}\rangle_h = -\tfrac{1}{2}\lambda_{\rm GS}\|\boldsymbol{\xi}\|_2^2 + O(\|\boldsymbol{\xi}\|_2^3),\\ \tfrac{1}{2}\big\langle \mathbf{v}-\mathbf{u}_{\rm GS}, \big(-\Delta_h+\mathbf{n}'(\mathbf{u}_{\rm GS})\big)(\mathbf{v}-\mathbf{u}_{\rm GS})\big\rangle_h &= \tfrac{1}{2}\big\langle \boldsymbol{\xi}, \big(-\Delta_h+\mathbf{n}'(\mathbf{u}_{\rm GS})\big)\boldsymbol{\xi}\big\rangle_h + O(\|\boldsymbol{\xi}\|_2^3),\end{aligned}$$

thus the Taylor expansion of f at $\boldsymbol{\xi}=0$ is

$$f(\boldsymbol{\xi}) = E_{\rm GS} - \tfrac{1}{2}\lambda_{\rm GS}\|\boldsymbol{\xi}\|_2^2 + \tfrac{1}{2}\big\langle \boldsymbol{\xi}, \big(-\Delta_h+\mathbf{n}'(\mathbf{u}_{\rm GS})\big)\boldsymbol{\xi}\big\rangle_h + O(\|\boldsymbol{\xi}\|_2^3) = E_{\rm GS} + \tfrac{1}{2}\langle \boldsymbol{\xi}, \mathcal{H}\boldsymbol{\xi}\rangle_h + O(\|\boldsymbol{\xi}\|_2^3),$$

where the last equality is (5.4). Here $\boldsymbol{\xi}$ ranges over the linear space $T_{\rm GS}$ and need not lie on $\mathcal{M}$, only the point $\mathbf{v}(\boldsymbol{\xi})$ does. Since $\mathbf{E}_h$ is a polynomial and $\mathbf{v}$ is smooth on $T_{\rm GS}$, f is a smooth function on $T_{\rm GS}$. Since f is defined on the inner product space $T_{\rm GS}$, its gradient at $\boldsymbol{\xi}$ is the unique vector $\nabla f(\boldsymbol{\xi})\in T_{\rm GS}$ with $\langle \nabla f(\boldsymbol{\xi}), \mathbf{w}\rangle_h = \frac{d}{dt}f(\boldsymbol{\xi}+t\mathbf{w})\big|_{t=0}$ for all $\mathbf{w}\in T_{\rm GS}$, and its Hessian at $\boldsymbol{\xi}$ is the h-self-adjoint operator $\nabla^2 f(\boldsymbol{\xi})$ on $T_{\rm GS}$ with $\langle \mathbf{w}, \nabla^2 f(\boldsymbol{\xi})\mathbf{w}\rangle_h = \frac{d^2}{dt^2}f(\boldsymbol{\xi}+t\mathbf{w})\big|_{t=0}$ for all $\mathbf{w}\in T_{\rm GS}$. By the uniqueness of the second-order Taylor polynomial of a smooth function, the expansion of f at $\boldsymbol{\xi}=0$ gives $\nabla f(0)=0$ and $\langle \mathbf{w}, \nabla^2 f(0)\mathbf{w}\rangle_h = \langle \mathbf{w}, \mathcal{H}\mathbf{w}\rangle_h$ for $\mathbf{w}\in T_{\rm GS}$, which is at least $\nu\|\mathbf{w}\|_2^2$ by the definition of ν applied to $\mathbf{w}/\|\mathbf{w}\|_2$ for $\mathbf{w}\neq 0$. Let $\mathbf{e}_1,\dots,\mathbf{e}_{N-1}$ be an $\langle\cdot,\cdot\rangle_h$-orthonormal basis of $T_{\rm GS}$. Expanding $\mathbf{w} = \sum_{i=1}^{N-1}\langle \mathbf{w}, \mathbf{e}_i\rangle_h\mathbf{e}_i$ in both arguments and using the Cauchy–Schwarz inequality for the sums over the index pairs (i,j), we have, for all $\boldsymbol{\xi},\mathbf{w}\in T_{\rm GS}$,

$$\begin{aligned}\big|\big\langle \mathbf{w}, \big(\nabla^2 f(\boldsymbol{\xi}) - \nabla^2 f(0)\big)\mathbf{w}\big\rangle_h\big| &= \Big|\sum_{i,j=1}^{N-1}\langle \mathbf{w}, \mathbf{e}_i\rangle_h\langle \mathbf{w}, \mathbf{e}_j\rangle_h\big\langle \mathbf{e}_i, \big(\nabla^2 f(\boldsymbol{\xi}) - \nabla^2 f(0)\big)\mathbf{e}_j\big\rangle_h\Big|\\ &\le \Big(\sum_{i,j=1}^{N-1}\big\langle \mathbf{e}_i, \big(\nabla^2 f(\boldsymbol{\xi}) - \nabla^2 f(0)\big)\mathbf{e}_j\big\rangle_h^2\Big)^{1/2}\Big(\sum_{i,j=1}^{N-1}\langle \mathbf{w}, \mathbf{e}_i\rangle_h^2\langle \mathbf{w}, \mathbf{e}_j\rangle_h^2\Big)^{1/2},\end{aligned}\tag{5.7}$$

where the last factor equals $\|\mathbf{w}\|_2^2$, since $\sum_{i,j=1}^{N-1}\langle \mathbf{w}, \mathbf{e}_i\rangle_h^2\langle \mathbf{w}, \mathbf{e}_j\rangle_h^2 = \big(\sum_{i=1}^{N-1}\langle \mathbf{w}, \mathbf{e}_i\rangle_h^2\big)^2 = \|\mathbf{w}\|_2^4$, and the other factor is the Frobenius norm of the matrix of $\nabla^2 f(\boldsymbol{\xi}) - \nabla^2 f(0)$ in this basis. Its entries are continuous functions of $\boldsymbol{\xi}$ that vanish at $\boldsymbol{\xi}=0$, thus there is $r>0$ such that the right side of (5.7) is at most $\frac{\nu}{2}\|\mathbf{w}\|_2^2$ for $\|\boldsymbol{\xi}\|_2\le r$. Then, for all $\boldsymbol{\xi},\mathbf{w}\in T_{\rm GS}$ with $\|\boldsymbol{\xi}\|_2\le r$,

$$\langle \mathbf{w}, \nabla^2 f(\boldsymbol{\xi})\mathbf{w}\rangle_h = \langle \mathbf{w}, \mathcal{H}\mathbf{w}\rangle_h + \big\langle \mathbf{w}, \big(\nabla^2 f(\boldsymbol{\xi}) - \nabla^2 f(0)\big)\mathbf{w}\big\rangle_h \ge \nu\|\mathbf{w}\|_2^2 - \tfrac{\nu}{2}\|\mathbf{w}\|_2^2 = \tfrac{\nu}{2}\|\mathbf{w}\|_2^2.$$

For such $\boldsymbol{\xi}$, Taylor's formula with integral remainder along the segment from 0 to $\boldsymbol{\xi}$ and $\nabla f(0)=0$ give

$$f(\boldsymbol{\xi}) - E_{\rm GS} = \int_0^1 (1-t)\langle \boldsymbol{\xi}, \nabla^2 f(t\boldsymbol{\xi})\boldsymbol{\xi}\rangle_h\,dt \ge \tfrac{\nu}{2}\|\boldsymbol{\xi}\|_2^2\int_0^1(1-t)\,dt = \tfrac{\nu}{4}\|\boldsymbol{\xi}\|_2^2,$$

and along the segment from $\boldsymbol{\xi}$ to 0, together with the Cauchy–Schwarz inequality and $as - \frac{\nu}{4}s^2 \le \frac{a^2}{\nu}$ for real a and s,

$$\begin{aligned} E_{\mathrm{GS}} - f(\boldsymbol{\xi}) &= -\langle \nabla f(\boldsymbol{\xi}), \boldsymbol{\xi}\rangle_h + \int_0^1 (1-t)\langle \boldsymbol{\xi}, \nabla^2 f((1-t)\boldsymbol{\xi})\boldsymbol{\xi}\rangle_h \, dt \\ &\ge -\|\nabla f(\boldsymbol{\xi})\|_2 \|\boldsymbol{\xi}\|_2 + \tfrac{\nu}{4}\|\boldsymbol{\xi}\|_2^2 \ge -\frac{1}{\nu}\|\nabla f(\boldsymbol{\xi})\|_2^2, \end{aligned}$$

Thus, for $\|\boldsymbol{\xi}\|_2 \le r$,

$$f(\boldsymbol{\xi}) - E_{\mathrm{GS}} \ge \tfrac{\nu}{4}\|\boldsymbol{\xi}\|_2^2, \qquad f(\boldsymbol{\xi}) - E_{\mathrm{GS}} \le \frac{1}{\nu}\|\nabla f(\boldsymbol{\xi})\|_2^2. \tag{5.8}$$

Next we estimate the gradient and the distance in the chart. For $\mathbf{w} \in T_{\mathrm{GS}}$, differentiating $\mathbf{v}(\boldsymbol{\xi})$ in the direction $\mathbf{w}$ gives

$$\mathbf{v}'(\boldsymbol{\xi})\mathbf{w} = \big(1 + \|\boldsymbol{\xi}\|_2^2\big)^{-1/2}\mathbf{w} - \big(1 + \|\boldsymbol{\xi}\|_2^2\big)^{-3/2}\langle \boldsymbol{\xi}, \mathbf{w}\rangle_h(\mathbf{u}_{\mathrm{GS}} + \boldsymbol{\xi}).$$

Differentiating $\|\mathbf{v}(\boldsymbol{\xi})\|_2^2 = 1$ gives $\langle \mathbf{v}(\boldsymbol{\xi}), \mathbf{v}'(\boldsymbol{\xi})\mathbf{w}\rangle_h = 0$. Expanding the square, with $\langle \mathbf{u}_{\mathrm{GS}} + \boldsymbol{\xi}, \mathbf{w}\rangle_h = \langle \boldsymbol{\xi}, \mathbf{w}\rangle_h$ and $\|\mathbf{u}_{\mathrm{GS}} + \boldsymbol{\xi}\|_2^2 = 1 + \|\boldsymbol{\xi}\|_2^2$, we have

$$\|\mathbf{v}'(\boldsymbol{\xi})\mathbf{w}\|_2^2 = \big(1 + \|\boldsymbol{\xi}\|_2^2\big)^{-1}\|\mathbf{w}\|_2^2 - 2\big(1 + \|\boldsymbol{\xi}\|_2^2\big)^{-2}\langle \boldsymbol{\xi}, \mathbf{w}\rangle_h^2 + \big(1 + \|\boldsymbol{\xi}\|_2^2\big)^{-2}\langle \boldsymbol{\xi}, \mathbf{w}\rangle_h^2 \le \|\mathbf{w}\|_2^2.$$

By the chain rule, $\nabla \mathbf{E}_h(\mathbf{v}) = \mathbf{g}(\mathbf{v}) + \lambda(\mathbf{v})\mathbf{v}$ from (3.3), $\langle \mathbf{v}(\boldsymbol{\xi}), \mathbf{v}'(\boldsymbol{\xi})\mathbf{w}\rangle_h = 0$, the Cauchy–Schwarz inequality and $\|\mathbf{v}'(\boldsymbol{\xi})\mathbf{w}\|_2 \le \|\mathbf{w}\|_2$,

$$\langle \nabla f(\boldsymbol{\xi}), \mathbf{w}\rangle_h = \langle \nabla \mathbf{E}_h(\mathbf{v}(\boldsymbol{\xi})), \mathbf{v}'(\boldsymbol{\xi})\mathbf{w}\rangle_h = \langle \mathbf{g}(\mathbf{v}(\boldsymbol{\xi})), \mathbf{v}'(\boldsymbol{\xi})\mathbf{w}\rangle_h \le \|\mathbf{g}(\mathbf{v}(\boldsymbol{\xi}))\|_2 \|\mathbf{w}\|_2,$$

and the choice $\mathbf{w} = \nabla f(\boldsymbol{\xi}) \in T_{\mathrm{GS}}$ gives $\|\nabla f(\boldsymbol{\xi})\|_2 \le \|\mathbf{g}(\mathbf{v}(\boldsymbol{\xi}))\|_2$. In addition, by $\|\mathbf{v}(\boldsymbol{\xi})\|_2 = \|\mathbf{u}_{\mathrm{GS}}\|_2 = 1$, $\langle \boldsymbol{\xi}, \mathbf{u}_{\mathrm{GS}}\rangle_h = 0$ and $(1+p)^{-1/2} \ge 1 - \frac{p}{2}$ for $p \ge 0$,

$$\|\mathbf{v}(\boldsymbol{\xi}) - \mathbf{u}_{\mathrm{GS}}\|_2^2 = 2 - 2\langle \mathbf{v}(\boldsymbol{\xi}), \mathbf{u}_{\mathrm{GS}}\rangle_h = 2 - 2\big(1 + \|\boldsymbol{\xi}\|_2^2\big)^{-1/2} \le \|\boldsymbol{\xi}\|_2^2.$$

Finally we prove the rate. By Theorem 3.4, $\mathbf{u}^k \to \mathbf{u}_{\mathrm{GS}}$, thus $\langle \mathbf{u}^k, \mathbf{u}_{\mathrm{GS}}\rangle_h \to 1$, and there is K such that, for all $k \ge K$, $\langle \mathbf{u}^k, \mathbf{u}_{\mathrm{GS}}\rangle_h > 0$ and $\langle \mathbf{u}^k, \mathbf{u}_{\mathrm{GS}}\rangle_h^{-2} - 1 \le r^2$, i.e., $\mathbf{u}^k = \mathbf{v}(\boldsymbol{\xi}_k)$ with $\boldsymbol{\xi}_k \in T_{\mathrm{GS}}$ and $\|\boldsymbol{\xi}_k\|_2 \le r$ by the chart. For such k, (5.8), $\|\nabla f(\boldsymbol{\xi}_k)\|_2 \le \big\|\mathbf{g}(\mathbf{u}^k)\big\|_2$, (5.3) and $\big\|\mathbf{u}^k - \mathbf{u}_{\mathrm{GS}}\big\|_2 \le \|\boldsymbol{\xi}_k\|_2$ give

$$\begin{aligned} \mathbf{E}_h(\mathbf{u}^k) - E_{\mathrm{GS}} &\le \frac{1}{\nu}\big\|\mathbf{g}(\mathbf{u}^k)\big\|_2^2 \le \frac{1 + \gamma\mu_{\max}}{\nu}\big\|\mathbf{g}(\mathbf{u}^k)\big\|_X^2, \\ \big\|\mathbf{u}^k - \mathbf{u}_{\mathrm{GS}}\big\|_2^2 &\le \|\boldsymbol{\xi}_k\|_2^2 \le \frac{4}{\nu}\big(\mathbf{E}_h(\mathbf{u}^k) - E_{\mathrm{GS}}\big). \end{aligned}$$

Plugging the first bound into the first inequality of (5.2), and using $\eta \le \frac{\gamma\nu}{4(1+\gamma\mu_{\max})}$ and $\eta \le \frac{1}{2}$ from (5.5), $\mathbf{E}_h(\mathbf{u}^k) - E_{\mathrm{GS}} \ge 0$ and $\mathcal{L}^k - E_{\mathrm{GS}} = \mathbf{E}_h(\mathbf{u}^k) - E_{\mathrm{GS}} + (R^k)^2/\gamma$ from (3.8), we obtain

$$\mathcal{L}^k - \mathcal{L}^{k+1} \ge \frac{\gamma\nu}{4(1 + \gamma\mu_{\max})}\big(\mathbf{E}_h(\mathbf{u}^k) - E_{\mathrm{GS}}\big) + \frac{1}{2}\frac{(R^k)^2}{\gamma} \ge \eta\big(\mathcal{L}^k - E_{\mathrm{GS}}\big),$$

and subtracting both sides from $\mathcal{L}^k - E_{\mathrm{GS}}$ gives the first inequality in (5.6). By induction, $\mathcal{L}^k - E_{\mathrm{GS}} \le (1-\eta)^{k-K}(\mathcal{L}^K - E_{\mathrm{GS}})$ for $k \ge K$, and adding $\frac{\nu}{4}\big\|\mathbf{u}^k - \mathbf{u}_{\mathrm{GS}}\big\|_2^2 \le \mathbf{E}_h(\mathbf{u}^k) - E_{\mathrm{GS}} \le \mathcal{L}^k - E_{\mathrm{GS}}$ to $\mathcal{L}^k - E_{\mathrm{GS}} = \mathbf{E}_h(\mathbf{u}^k) - E_{\mathrm{GS}} + (R^k)^2/\gamma$ gives the second inequality in (5.6). $\square$

6. Numerical Experiments

In this section, we test two splitting schemes (1.6) and (1.7) numerically, and compare them with five other efficient iterative methods: the H^1 gradient flow and the Riemannian conjugate gradient method in the same H^1 metric, the preconditioned gradient (PG) and preconditioned conjugate gradient (PCG) methods in [4], and the energy-adaptive a_u-flow [24]. Two spatial discretizations of $\Omega = [-8, 8]^3$ will be used. The first one is the second-order finite difference scheme, satisfying Assumption 2.1, on a uniform grid with n^3 interior points, so that $\mathbb{M} = h^3\mathbb{I}$ with $h = 16/(n + 1)$, at $n = 799$. The second discretization is the Q^k spectral element method with the diagonal mass matrix of Gauss–Lobatto quadrature [29, 12]. In numerical tests below, we use the Q^{20} spectral element method on 50 cells per direction, so that the 3D problem has 999^3 unknowns. The Q^{20} scheme is not monotone, so these runs lie outside the hypotheses of Theorem 3.4. By considering such a discretization, we would like to show that the performance belongs to the splitting schemes and not uniquely to one discretization.

In every experiment $\beta = 1600$, and every method built from $\mathbb{G}_\alpha := (\alpha\mathbb{I} - \Delta_h)^{-1}$ uses the shift $\alpha = 20$. All methods solve their linear systems by the same simple solver in [29, 30]. All tests were run in Python with JAX on one NVIDIA GH200 GPU card, in double precision.

6.1. The methods.

H^1 flow with line search is the discrete H^1 gradient flow [26, 15, 12], that is, Riemannian gradient descent on $\mathcal{M}$ for the shifted H^1 metric $\langle \mathbf{v}, \mathbf{w}\rangle_\alpha := \langle \mathbf{v}, (\alpha\mathbb{I} - \Delta_h)\mathbf{w}\rangle_h$. The Riemannian gradient of $\mathbf{E}_h$ at $\mathbf{u} \in \mathcal{M}$ for this metric is

$$\mathbf{g}_\alpha(\mathbf{u}) := \mathbb{G}_\alpha \nabla \mathbf{E}_h(\mathbf{u}) - \frac{\langle \mathbb{G}_\alpha \nabla \mathbf{E}_h(\mathbf{u}), \mathbf{u}\rangle_h}{\langle \mathbb{G}_\alpha \mathbf{u}, \mathbf{u}\rangle_h}\, \mathbb{G}_\alpha \mathbf{u}, \tag{6.1}$$

the $\langle\cdot,\cdot\rangle_\alpha$-orthogonal projection of $\mathbb{G}_\alpha \nabla \mathbf{E}_h(\mathbf{u})$ onto the tangent plane of $\mathcal{M}$ at $\mathbf{u}$, and one step is

$$\mathbf{u}^{k+1} = \frac{\mathbf{u}^k - \tau_k\, \mathbf{g}_\alpha(\mathbf{u}^k)}{\|\mathbf{u}^k - \tau_k\, \mathbf{g}_\alpha(\mathbf{u}^k)\|_2}, \tag{6.2}$$

which applies $\mathbb{G}_\alpha$ twice, to $\nabla \mathbf{E}_h(\mathbf{u}^k)$ and to $\mathbf{u}^k$. The step size τ_k is chosen at every iteration by minimizing the energy of the normalized update.

Riemannian CG is the Riemannian conjugate gradient method of Danaila and Protas [16] for the same metric: the search direction at $\mathbf{u}^k$ is $-\mathbf{g}_\alpha(\mathbf{u}^k)$ plus the previous direction, transported to the tangent plane at $\mathbf{u}^k$ by the projection in (6.1) and multiplied by the Fletcher–Reeves ratio $\langle \mathbf{g}_\alpha(\mathbf{u}^k), \mathbf{g}_\alpha(\mathbf{u}^k)\rangle_\alpha / \langle \mathbf{g}_\alpha(\mathbf{u}^{k-1}), \mathbf{g}_\alpha(\mathbf{u}^{k-1})\rangle_\alpha$, and the step size comes from an Armijo search started at the one-dimensional minimizer.

Shifted Laplacian preconditioned GF and *CG* are the PG and PCG methods of [4] with $\mathbb{G}_\alpha$ as the preconditioner: the same shifted Laplacian that defines the metric above, used as a preconditioner rather than as a metric, so the two families differ in how they use the operator and not in which one they invert. Both take the projected gradient $\mathbf{g}(\mathbf{u}^k)$ of (3.3) as the residual, with directions $\mathbf{d}^k = -\mathbb{G}_\alpha\, \mathbf{g}(\mathbf{u}^k)$ for PG and

$$\mathbf{d}^k = -\mathbb{G}_\alpha\, \mathbf{g}(\mathbf{u}^k) + \frac{\langle \mathbf{g}(\mathbf{u}^k), \mathbb{G}_\alpha \mathbf{g}(\mathbf{u}^k)\rangle_h}{\langle \mathbf{g}(\mathbf{u}^{k-1}), \mathbb{G}_\alpha \mathbf{g}(\mathbf{u}^{k-1})\rangle_h}\, \mathbf{p}^{k-1}$$

for PCG. Both project the direction onto the tangent plane, $\mathbf{p}^k = \mathcal{P}_{\mathbf{u}^k}\mathbf{d}^k$, and update along the great circle through $\mathbf{u}^k$,

$$\mathbf{u}^{k+1} = \cos(\theta_k)\, \mathbf{u}^k + \sin(\theta_k)\, \frac{\mathbf{p}^k}{\|\mathbf{p}^k\|_2}, \tag{6.3}$$

with θ_k minimizing $\mathbf{E}_h(\mathbf{u}^{k+1})$ over $[-\pi, \pi]$. Each iteration applies $\mathbb{G}_\alpha$ once.

a_u-flow is the energy-adaptive Riemannian gradient flow of [24], for the metric induced by $A_\mathbf{u}$ itself, $\langle \mathbf{v}, \mathbf{w} \rangle_{a_u} = \langle \mathbf{v}, A_\mathbf{u}\mathbf{w} \rangle_h$. Because $\nabla \mathbf{E}_h(\mathbf{u}) = A_\mathbf{u}\mathbf{u}$, applying $A_\mathbf{u}^{-1}$ to the gradient costs nothing, so one solve per iteration suffices rather than two, as recorded in Table 1: with $\mathbf{w}^k$ the solution of $A_{\mathbf{u}^k}\mathbf{w}^k = \mathbf{u}^k$ and $\rho_k = \langle \mathbf{u}^k, \mathbf{u}^k \rangle_h / \langle \mathbf{u}^k, \mathbf{w}^k \rangle_h$, the step is

$$\mathbf{u}^{k+1} = \frac{\mathbf{u}^k + \tau_k(\rho_k \mathbf{w}^k - \mathbf{u}^k)}{\|\mathbf{u}^k + \tau_k(\rho_k \mathbf{w}^k - \mathbf{u}^k)\|_2}, \tag{6.4}$$

which at $\tau_k = 1$ is a normalized inverse iteration in the a_u metric. The matrix $A_{\mathbf{u}^k}$ moves with the iterate and is not separable, so this is the one linear system in the comparison that cannot be diagonalized. It is solved by conjugate gradients preconditioned by $(-\Delta_h + \mathbb{V}_1)^{-1}$ as in [30], stopped at relative residual 10^{-8}. The step size is taken from [24, Remark 4.3] rather than fixed at 1: along the update ray the energy of the normalized iterate is a ratio of polynomials in τ whose eleven coefficients are inner products of $\mathbf{u}^k$ and $\rho_k \mathbf{w}^k$, gathered once per step, after which every trial τ is a scalar evaluation and the minimization over $(0, 2)$ needs no further solve. Without this rule the a_u-flow would be the only method here carrying an untuned step size: at $n = 199$ it converges in 54 iterations against 77 at $\tau_k = 1$, and against 69 at the best constant τ found by a sweep.

Davis–Yin I is the scheme (1.6), whose single solve per iteration inverts $\mathbb{I} - \gamma_k \Delta_h$, with

$$L_k := \max_i \bigl(V_i + 3\beta(u_i^k)^2\bigr), \tag{6.5}$$

the Lipschitz constant at the current iterate of ∇H, where $H(\mathbf{u}) = \frac{1}{2}\langle \mathbb{V}\mathbf{u}, \mathbf{u} \rangle_h + \frac{\beta}{4}\|\mathbf{u}\|_4^4$ is the explicitly treated part of the splitting in Appendix B, since $\nabla^2 H(\mathbf{u}) = \operatorname{diag}\bigl(V_i + 3\beta u_i^2\bigr)$. The choice $\gamma_k = 1/L_k$ is already larger than the threshold $\gamma_{\max}$ of Theorem 3.4, which is what the proof requires rather than what the method tolerates. Even $1/L_k$ is conservative by about a factor of two, so in practice we use $\gamma_k = c/L_k$. Different constants give different performance, and the values used here were found by tuning on coarse grids: for the three examples of §6.2, $c = 1.6$ on the lattice from the constant initial guess, 1.7 on the lattice from the linear one and 2.1 on the stirrer. *Davis–Yin II* is the scheme (1.7), the same iteration with $-\Delta_h + \mathbb{V}_1$ inside the resolvent and with $L_k = \max_i(V_{2,i} + 3\beta(u_i^k)^2)$. Its constants, tuned the same way, are smaller, 1.4 on the lattice from either initial guess and 1.2 on the stirrer, since a smaller L_k comes with a tighter limit on c and only the product $\gamma_k = c/L_k$ matters.

Five of the seven methods choose their step size by a one-dimensional minimization and the two Davis–Yin schemes do not, as Table 2 records. Every such search is evaluated in closed form, the energy along the search direction being a ratio of polynomials in the step, so that no line search costs a linear solve. Table 3 records what one iteration of each method costs, in elliptic solves and in memory.

6.2. **The two potentials and the three examples.** The figures below plot the energy error $|\mathbf{E}_h(\mathbf{u}^k) - E^*|$ against the iteration count and against the wall-clock time. The reference energy E^* is generated numerically on the same grid by running any one of the methods for enough iterations, well past the accuracy that the compared runs reach, so that no plotted curve is limited by it.

The first potential is an optical lattice superposed on a harmonic trap,

$$V_{\mathrm{lat}}(\boldsymbol{x}) = |\boldsymbol{x}|^2 + 100 \sum_{j=1}^{3} \sin^2\Bigl(\frac{\pi x_j}{4}\Bigr), \tag{6.6}$$

method	step size	one-dimensional minimization
H^1 flow with line search	τ_k minimizing $\mathbf{E}_h$ in (6.2)	along the ray, over $\tau \in (0, 5]$
Riemannian CG	Armijo from that minimizer	along the ray, then a backtrack
Shifted Laplacian preconditioned GF	θ_k minimizing $\mathbf{E}_h$ in (6.3)	along the great circle, $[-\pi, \pi]$
Shifted Laplacian preconditioned CG	as above	along the great circle
DYS I	$\gamma_k = c/L_k,\ L_k = \max_i(V_i + 3\beta(u_i^k)^2)$	none
DYS II	$\gamma_k = c/L_k,\ L_k = \max_i(V_{2,i} + 3\beta(u_i^k)^2)$	none
a_u-flow	τ_k minimizing $\mathbf{E}_h$ in (6.4)	along the update ray, over $(0, 2)$

TABLE 2. The step size rules used in numerical tests.

method	vectors carried across an iteration	elliptic solves per iteration, and of what	memory cost per iteration of the simple efficient implementation used in this paper, on a 999^3 grid
DYS I (1.6)	$2N$	$1 \times (\mathbb{I} - \gamma_k \Delta_h)^{-1}$	$8\,n^3 + 2n^2$: 63.8 GB
DYS II (1.7)	$2N$	$1 \times (\mathbb{I} - \gamma_k \Delta_h + \gamma_k \mathbb{V}_1)^{-1}$	$8\,n^3 + 2n^2$: 63.8 GB
Shifted Laplacian precond. CG	$2N$	$1 \times (\alpha\mathbb{I} - \Delta_h)^{-1}$	$7\,n^3 + 2n^2$: 55.8 GB
Shifted Laplacian precond. GF	$2N$	$1 \times (\alpha\mathbb{I} - \Delta_h)^{-1}$	$9\,n^3 + 2n^2$: 71.8 GB
H^1 flow with line search	$2N$	$2 \times (\alpha\mathbb{I} - \Delta_h)^{-1}$	$9\,n^3 + 2n^2$: 71.8 GB
Riemannian CG	$2N$	$2 \times (\alpha\mathbb{I} - \Delta_h)^{-1}$	$8\,n^3 + 2n^2$: 63.8 GB
a_u-flow	$1N$	$1 \times A_{\mathbf{u}^k}^{-1}$	$10\,n^3 + 2n^2$: 79.8 GB

TABLE 3. What one iteration costs, with $N = n^3$ degrees of freedom. The first two columns belong to the schemes and hold for any implementation. The last is what this one reaches, measured one method per process on one GH200 in double precision. The $2n^2$ is the pair of $n \times n$ factors of the per-axis eigendecomposition, which is needed in the simple GPU implementation of Poisson solver [29, 30].

which is additively separable. The second is an anisotropic harmonic trap stirred by a Gaussian beam directed along x_3,

$$V_{\text{stir}}(\boldsymbol{x}) = x_1^2 + x_2^2 + \omega_z^2 x_3^2 + 2w_0\, e^{-\delta\left((x_1-r_0)^2+x_2^2\right)}, \qquad w_0 = 30,\ \delta = 1,\ r_0 = 1,\ \omega_z = 2, \tag{6.7}$$

a standard model of a laser stirrer in a condensate [6]. The beam couples x_1 and x_2, so V_{stir} is not separable. Writing $V = V_1 + V_2$ with V_1 the separable part, the Gaussian V_2 is a rank-one outer product in (x_1, x_2). The scheme (1.6) treats V explicitly and is indifferent to this distinction. The scheme (1.7) and the a_u-flow are not, the first inverting $-\Delta_h + \mathbb{V}_1$ and the second using it as a preconditioner. Both need V_1 to be separable, and $-\Delta_h + \mathbb{V}_1$ is then inverted by the fast diagonalization of [30]. Figure 1 shows both traps and the ground states they support.

Example 6.1 (Optical lattice, constant initial guess). The potential is (6.6) and every method starts from the normalized constant vector, which is positive as Theorem 3.4 requires. Figure 2 reports the comparison, at $n = 799$ with finite differences and at $n = 999$ with spectral elements. The practical computational efficiency of the two Davis–Yin splitting schemes is comparable to that of the CG methods, even though they need more iterations.

Example 6.2 (Stirrer, constant initial guess). The potential is (6.7), which is not separable, and the initial guess is again the normalized constant vector. Figure 3 reports the comparison on the same two grids. The stirring beam is explicit in (1.6), and its peak sets L_k and with it the step size, so

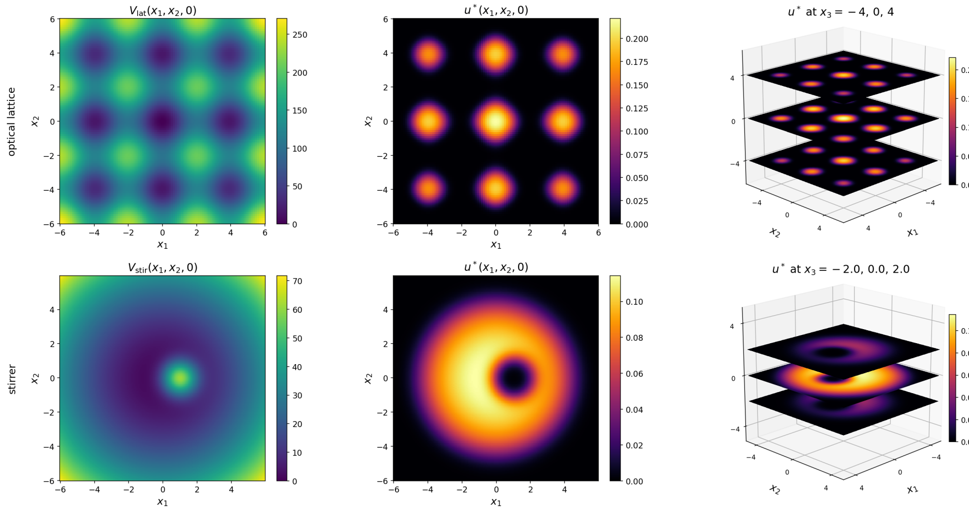


FIGURE 1. The two potentials and their ground states: the optical lattice (6.6) on top, the stirrer (6.7) below. Left is the trap on the plane $x_3 = 0$, middle the ground state on the same plane, right the ground state on three planes.

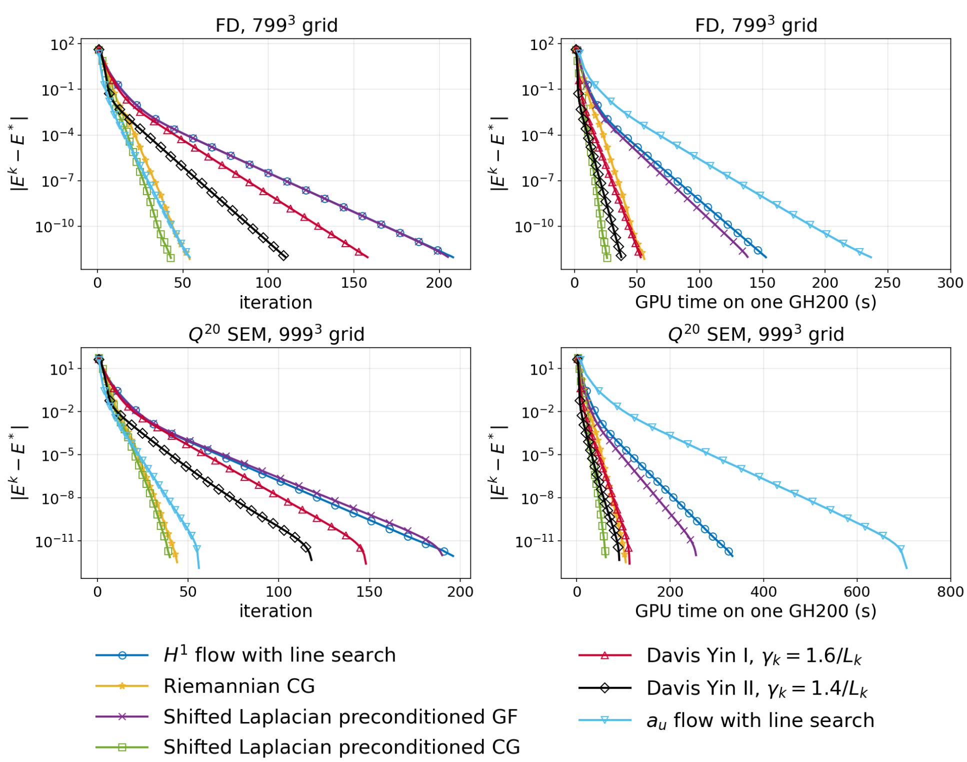


FIGURE 2. Example 6.1, the optical lattice from the constant initial guess: second-order finite differences (FD) on a 799^3 grid on top, Q^{20} spectral elements (SEM) on a 999^3 grid below.

here (1.6) is slower than the CG methods. The scheme (1.7) keeps only the beam explicit, and is the fastest of the seven methods in wall-clock time.

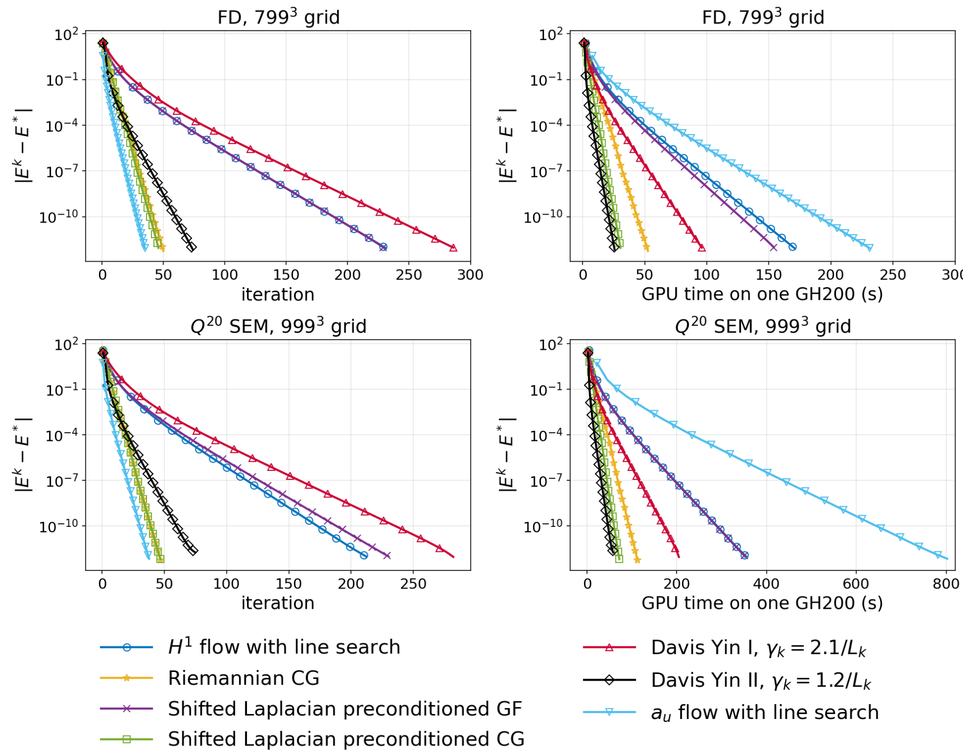
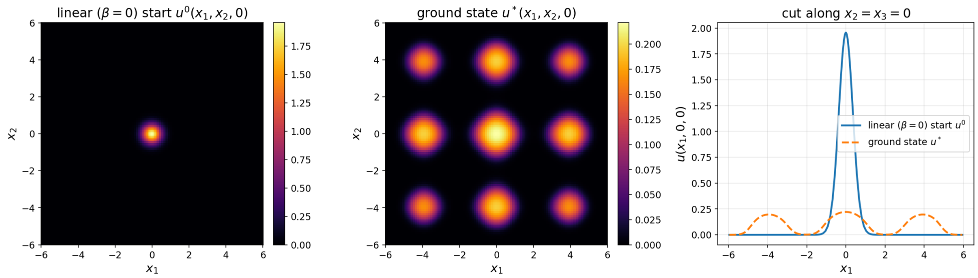


FIGURE 3. Example 6.2, the stirrer potential, in the panels of Figure 2. DYS II is the fastest of the seven methods here.

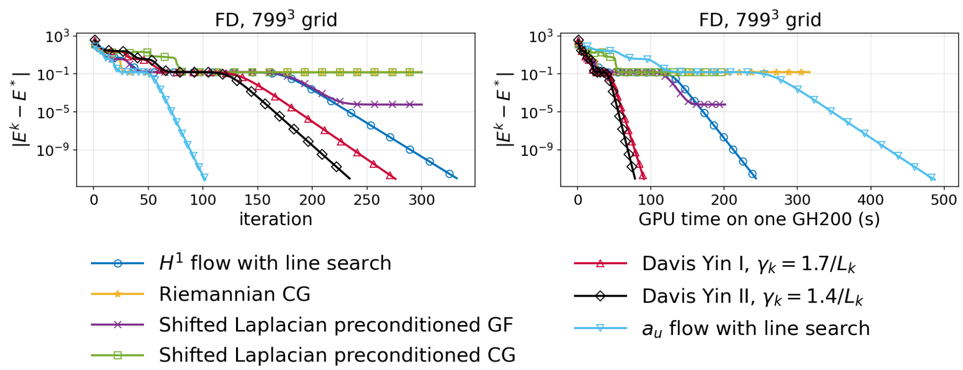
Example 6.3 (Optical lattice, linear ground state initial guess). The potential is (6.6) and every method starts from the ground state of the linear problem, the eigenvector of $-\Delta_h + \mathbb{V}$ belonging to its smallest eigenvalue, computed by shifted inverse iteration. This is a natural and widely used initialization for smaller β, but at $\beta = 1600$ its energy exceeds E^* by a factor of about sixteen. Figure 4 shows the initial guess and the comparison. Unlike the conjugate gradient methods, which are sensitive to this poor initial guess and stall within the first few hundred iterations at states that are not the ground state, the schemes (1.6) and (1.7) converge very efficiently.

7. Concluding remarks

We have proved global convergence of the semi-implicit Davis–Yin splitting scheme (1.6) for the discrete defocusing Gross–Pitaevskii ground state problem: for the monotone discretizations of Section 2, for every positive normalized initial vector, and for every constant step size below the explicit threshold of Theorem 3.4, the normalized iterates converge to the unique positive discrete ground state, with no spectral gap assumption and no requirement that the initial guess be close to it. In the numerical tests of Section 6, the simple step size rule $\gamma_k = c/L_k$ of (6.5), which costs no line search and no inner iteration, makes the schemes (1.6) and (1.7) efficient in practice: on the



(a) The linear ground state ($\beta = 0$), and the ground state for $\beta = 1600$ for the lattice potential.



(b) The comparison for second order finite difference on a 799^3 grid.

FIGURE 4. At $\beta = 1600$ the ground state of the linear problem is a poor initial guess, which is used in Example 6.3 only to test how much methods depend on a good initial guess. The CG methods perform poorly, while the DYS schemes are superior in computational time.

examples where every method converges, their wall-clock performance is comparable to that of the conjugate gradient methods, and they are much more robust with respect to the choice of the initial guess.

Appendix A. The monotone spatial discretizations

This appendix collects the construction of the two discretizations of Section 2 and the monotonicity properties they enjoy, following [12].

A.1. **Finite element method with quadrature.** Let Ω_h be a mesh of Ω, which is either

(1) a uniform rectangular mesh, on which we use the space of continuous piecewise Q^1 (multi-linear) polynomials, or
(2) an unstructured simplicial mesh, e.g., a triangular mesh for $d = 2$, on which we use the space of continuous piecewise P^1 (linear) polynomials.

In both cases the finite element space with the homogeneous Dirichlet boundary condition is

$$V_0^h = \{v_h \in C(\Omega) : \ v_h|_{\partial\Omega} = 0, \ v_h|_e \in Q^1(e) \text{ (resp. } P^1(e)), \ \forall\, e \in \Omega_h\} \subset H_0^1(\Omega).$$

In practice the integrals defining the scheme are evaluated by quadrature. For the Q^1 method we use the 2-point Gauss–Lobatto (i.e., trapezoidal) rule in each dimension on each cell, and for the P^1 method the vertex rule $\int_e f \,\mathrm{d}\boldsymbol{x} \approx \frac{|e|}{d+1}\sum_{\boldsymbol{x}_v \in e} f(\boldsymbol{x}_v)$ on each simplex, i.e., mass lumping. (For P^1 elements the stiffness integrand $\nabla u_h \cdot \nabla v_h$ is piecewise constant, so it is integrated exactly, and only the mass and potential terms are lumped.) Let $\langle \cdot, \cdot \rangle$ denote the resulting quadrature approximation of $(\cdot,\cdot)$. The finite element method with quadrature for (1.4) is to seek $\lambda_h \in \mathbb{R}$ and $u_h \in V_0^h$ satisfying

$$\langle \nabla u_h, \nabla v_h \rangle + \langle V u_h, v_h \rangle + \beta \langle |u_h|^2 u_h, v_h \rangle = \lambda_h \langle u_h, v_h \rangle, \qquad \forall\, v_h \in V_0^h, \tag{A.1}$$

with the corresponding discrete energy

$$E_h(u_h) = \frac{1}{2}\langle \nabla u_h, \nabla u_h \rangle + \frac{1}{2}\langle V u_h, u_h \rangle + \frac{\beta}{4}\langle u_h^2, u_h^2 \rangle, \tag{A.2}$$

minimized under the normalization constraint $\langle u_h, u_h \rangle = 1$. On a uniform rectangular mesh the Q^1 scheme with quadrature coincides with the classical second-order central finite difference scheme, see [12, Appendix] for the explicit expressions of both schemes. Using such quadrature does not degrade the standard *a priori* error estimates of the finite element method [14]. For the convergence of finite element approximations of the nonlinear eigenvalue problem (1.3), see [9, 12].

A.2. **Matrix–vector form.** With the notation of Section 2, write $u_i = u_h(\boldsymbol{x}_i)$ and $u_h = \sum_{i=1}^N u_i \phi_i$ for $u_h \in V_0^h$. Then

$$\langle V u_h, v_h \rangle = \sum_{i=1}^{N} w_i V_i u_i v_i = \mathbf{v}^\top \mathbb{M}\mathbb{V}\mathbf{u}, \qquad \langle \nabla u_h, \nabla v_h \rangle = \mathbf{v}^\top \mathbb{S}\mathbf{u}, \tag{A.3}$$

so that (A.1) is exactly (2.1) and (A.2) is exactly (2.4). The stiffness matrix $\mathbb{S}$ is symmetric positive definite, and $\mathbb{M}$ is diagonal with positive entries.

A.3. **Monotonicity.** A matrix A is called *monotone* if $A^{-1} \geq 0$ entrywise. A symmetric positive definite matrix with nonpositive off-diagonal entries is a (symmetric) M-matrix, also called a Stieltjes matrix, and is monotone, see [35, 32]. Both discretizations above are monotone in the following sense.

(1) For the Q^1 scheme on a uniform rectangular mesh (equivalently, the second-order finite difference scheme), the stiffness matrix $\mathbb{S}$ has nonpositive off-diagonal entries, which is classical.

(2) For the P^1 scheme on a simplicial mesh $\Omega_h \subset \mathbb{R}^d$, let E denote an interior edge connecting the vertices $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$, let κ_E^T be the $(d-2)$-dimensional simplex opposite to E in a simplex $T \supset E$, and let θ_E^T be the dihedral angle between the two faces of T containing E. Then $\mathbb{S}_{ij} = -\sum_{T \supset E} \frac{1}{d(d-1)} |\kappa_E^T| \cot\theta_E^T$, so the off-diagonal entries of $\mathbb{S}$ are nonpositive if and only if the mesh satisfies

$$\sum_{T \supset E} \frac{1}{d(d-1)} |\kappa_E^T| \cot\theta_E^T \ \geq\ 0 \qquad \text{for every interior edge } E, \tag{A.4}$$

see [37, Lemma 2.1]. For $d = 2$ the condition (A.4) reduces to $\cot\theta_E^{T_1} + \cot\theta_E^{T_2} \geq 0$, i.e., $\theta_E^{T_1} + \theta_E^{T_2} \leq \pi$ for the two angles opposite to each interior edge: this is exactly the defining property of a *Delaunay triangulation*, which is more general and more practical than a nonobtuse triangulation. A mesh in which every simplex is nonobtuse satisfies (A.4) in any dimension.

In both cases the adjacency graph of $\mathbb{S}$ contains the edge–vertex connectivity graph of the mesh, and we assume it is connected, i.e., $\mathbb{S}$ is *irreducible*. This holds automatically for the finite difference scheme, and for the P^1 scheme whenever (A.4) holds strictly on enough edges, e.g., for a strictly Delaunay triangulation, see [27]. This is the connectivity required in Assumption 2.1, and with it items (i) and (ii) of Lemma 2.2 hold for both schemes.

Consequently $\mathbb{M}A_{\mathbf{u}} = \mathbb{S}+\mathbb{M}\mathbb{V}+\beta\mathbb{M}\,\mathrm{diag}(\mathbf{u}^2)$ is an irreducible symmetric M-matrix for every $\mathbf{u}$, so $A_{\mathbf{u}}$ is monotone and, by the Perron–Frobenius theorem, the smallest eigenvalue of $A_{\mathbf{u}}$ is simple and its eigenvector is the only nonnegative one, up to normalization [12, Section 3]. This yields the two facts quoted at the end of Section 2. First, as shown in [12, Section 3], the discrete energy inherits the hidden convexity of the continuous problem, being convex as a function of the discrete density, so that (2.5) has a minimizer that is unique up to sign, and its positive representative is the discrete ground state $\mathbf{u}_{\mathrm{GS}}$. Second, if $\mathbf{u} \geq 0$ is a critical point of (2.5), then $A_{\mathbf{u}}\mathbf{u} = \lambda_h \mathbf{u}$ with $\mathbf{u} \geq 0$ and $\mathbf{u} \neq 0$, so by Perron–Frobenius $\mathbf{u}$ must be the eigenvector belonging to the smallest eigenvalue of $A_{\mathbf{u}}$. In particular $\mathbf{u} > 0$, and the convexity in the density variables then forces $\mathbf{u} = \mathbf{u}_{\mathrm{GS}}$. The same conclusions were obtained for mass-lumped P^1 finite elements on unstructured meshes in [21] by the inverse positivity of irreducible M-matrices and a discrete Picone inequality, avoiding the explicit use of the Perron–Frobenius theorem.

Appendix B. The Davis–Yin splitting

This appendix recalls the Davis–Yin splitting and derives the two schemes (1.6) and (1.7) of Section 1.3. Throughout, the proximal operator of a function φ with parameter $\gamma > 0$ is taken with respect to the discrete inner product (2.3),

$$\mathrm{prox}_{\varphi}^{\gamma}(\mathbf{v}) := \mathrm{argmin}_{\mathbf{w}\in\mathbb{R}^N}\ \ \varphi(\mathbf{w}) + \frac{1}{2\gamma}\left\|\mathbf{w}-\mathbf{v}\right\|_2^2. \tag{B.1}$$

B.1. **The three-term splitting.** The discrete ground state problem (2.5) is the unconstrained minimization

$$\min_{\mathbf{u}\in\mathbb{R}^N}\ F(\mathbf{u}) + G(\mathbf{u}) + H(\mathbf{u}), \tag{B.2}$$

of the three functions

$$F(\mathbf{u}) := \tfrac{1}{2}\langle -\Delta_h \mathbf{u}, \mathbf{u}\rangle_h, \qquad G(\mathbf{u}) := \iota_{\mathcal{M}}(\mathbf{u}), \qquad H(\mathbf{u}) := \tfrac{1}{2}\langle \mathbb{V}\mathbf{u}, \mathbf{u}\rangle_h + \tfrac{\beta}{4}\left\|\mathbf{u}\right\|_4^4, \tag{B.3}$$

where F is the discrete Dirichlet (kinetic) energy, H collects the potential and interaction energies, so that $F + H = \mathbf{E}_h$ is the discrete energy (2.4), and $\iota_{\mathcal{M}}$ is the indicator function of $\mathcal{M}$, equal to 0 on $\mathcal{M}$ and $+\infty$ elsewhere, which enforces the normalization constraint.

For F, the first-order condition for the minimization in (B.1) is $\gamma(-\Delta_h\mathbf{w}) + \mathbf{w} - \mathbf{v} = 0$, so that

$$\mathrm{prox}_{F}^{\gamma}(\mathbf{v}) = (\mathbb{I} - \gamma\Delta_h)^{-1}\mathbf{v} = (\mathbb{M} + \gamma\mathbb{S})^{-1}\mathbb{M}\mathbf{v}, \tag{B.4}$$

i.e., one solve of a linear system with the symmetric positive definite matrix $\mathbb{M} + \gamma\mathbb{S}$. For G, the proximal operator of the indicator function of a set is the projection onto that set, and the projection onto $\mathcal{M}$ is the normalization

$$\mathrm{prox}_{G}^{\gamma}(\mathbf{v}) = \frac{\mathbf{v}}{\left\|\mathbf{v}\right\|_2}, \qquad \mathbf{v} \neq 0, \tag{B.5}$$

which is independent of γ and costs no linear solve. Finally H is smooth, with

$$\nabla H(\mathbf{u}) = \mathbb{V}\mathbf{u} + \beta\mathbf{u}^3, \tag{B.6}$$

evaluated entrywise at no cost.

B.2. **The Davis–Yin iteration.** For a sum of three functions as in (B.2), in which two terms are handled by their proximal operators and the third by its gradient, the Davis–Yin splitting [17] reads

$$\begin{cases} \mathbf{u}^k = \mathrm{prox}_G^\gamma(\mathbf{z}^k), \\ \mathbf{x}^k = \mathrm{prox}_F^\gamma\big(2\mathbf{u}^k - \mathbf{z}^k - \gamma\nabla H(\mathbf{u}^k)\big), \\ \mathbf{z}^{k+1} = \mathbf{z}^k + \mathbf{x}^k - \mathbf{u}^k. \end{cases} \tag{B.7}$$

It contains two classical schemes as special cases: for $H \equiv 0$ it reduces to the Douglas–Rachford splitting, and for $G \equiv 0$ to the forward–backward splitting, which is why it is also known as the forward–Douglas–Rachford splitting [33]. For *convex* F, G, H with ∇H Lipschitz continuous with constant L, the iteration (B.7) converges for any step size $\gamma < 2/L$ [17]. Substituting (B.4)–(B.6) into (B.7) gives the scheme (1.6) analyzed in this paper.

An alternative splitting moves part of the potential into F. Write $V = V_1 + V_2$ with $V_1, V_2 \geq 0$, and let $\mathbb{V}_1$ and $\mathbb{V}_2$ be the corresponding diagonal matrices, so that $\mathbb{V} = \mathbb{V}_1 + \mathbb{V}_2$. In place of (B.3) take

$$F(\mathbf{u}) := \tfrac{1}{2}\langle(-\Delta_h + \mathbb{V}_1)\mathbf{u}, \mathbf{u}\rangle_h, \qquad G(\mathbf{u}) := \iota_{\mathcal{M}}(\mathbf{u}), \qquad H(\mathbf{u}) := \tfrac{1}{2}\langle\mathbb{V}_2\mathbf{u}, \mathbf{u}\rangle_h + \tfrac{\beta}{4}\left\|\mathbf{u}\right\|_4^4, \tag{B.8}$$

so that again $F + H = \mathbf{E}_h$. The first-order condition for the minimization in (B.1) now reads $\gamma(-\Delta_h + \mathbb{V}_1)\mathbf{w} + \mathbf{w} - \mathbf{v} = 0$, so

$$\mathrm{prox}_F^\gamma(\mathbf{v}) = (\mathbb{I} - \gamma\Delta_h + \gamma\mathbb{V}_1)^{-1}\mathbf{v} = (\mathbb{M} + \gamma\mathbb{S} + \gamma\mathbb{M}\mathbb{V}_1)^{-1}\mathbb{M}\mathbf{v}, \tag{B.9}$$

while prox_G^γ is unchanged and $\nabla H(\mathbf{u}) = \mathbb{V}_2\mathbf{u} + \beta\mathbf{u}^3$. Substituting these into the iteration (B.7) gives the scheme (1.7). The matrix $\mathbb{M} + \gamma\mathbb{S} + \gamma\mathbb{M}\mathbb{V}_1$ differs from $\mathbb{M} + \gamma\mathbb{S}$ by a nonnegative diagonal matrix, so it is still an irreducible Stieltjes matrix and its inverse has strictly positive entries, by the argument of Lemma 2.2(ii).

Acknowledgments

This work used DeltaAI at the National Center for Supercomputing Applications (NCSA) through allocation MTH260013 from the Advanced Cyberinfrastructure Coordination Ecosystem: Services & Support (ACCESS) program, which is supported by U.S. National Science Foundation grants #2138259, #2138286, #2138307, #2137603, and #2138296.

Declaration of AI-assisted technologies

The authors have used generative AI tools such as ChatGPT and Claude to assist mathematical discussions, algorithm implementations and drafting the manuscript. After using AI tools, the authors have carefully reviewed, edited and polished the manuscript, and take full responsibility for the content.

References

[1] Robert Altmann, Patrick Henning, and Daniel Peterseim, *The J-method for the Gross-Pitaevskii eigenvalue problem*, Numerische Mathematik **148** (2021), no. 3, 575–610.

[2] Robert Altmann, Daniel Peterseim, and Tatjana Stykel, *Energy-adaptive Riemannian optimization on the Stiefel manifold*, ESAIM: Mathematical Modelling and Numerical Analysis **56** (2022), no. 5, 1629–1653.

[3] Robert Altmann, Daniel Peterseim, and Tatjana Stykel, *Riemannian Newton methods for energy minimization problems of Kohn–Sham type*, Journal of Scientific Computing **101** (2024), no. 1, Paper No. 6.

[4] Xavier Antoine, Antoine Levitt, and Qinglin Tang, *Efficient spectral computation of the stationary states of rotating Bose–Einstein condensates by preconditioned nonlinear conjugate gradient methods*, Journal of Computational Physics **343** (2017), 92–109.

[5] Weizhu Bao and Yongyong Cai, *Mathematical theory and numerical methods for Bose-Einstein condensation*, Kinetic & Related Models **6** (2013), no. 1, 1–135.

[6] Weizhu Bao and Qiang Du, *Computing the ground state solution of Bose-Einstein condensates by a normalized gradient flow*, SIAM Journal on Scientific Computing **25** (2004), no. 5, 1674–1697.

[7] Abraham Berman and Robert J. Plemmons, *Nonnegative matrices in the mathematical sciences*, Classics in Applied Mathematics 9, SIAM, Philadelphia, 1994.

[8] Eric Cancès, *SCF algorithms for HF electronic calculations*, Mathematical Models and Methods for Ab Initio Quantum Chemistry (Mireille Defranceschi and Claude Le Bris, eds.), Lecture Notes in Chemistry, vol. 74, Springer, Berlin, 2000, pp. 17–43.

[9] Eric Cancès, Rachida Chakir, and Yvon Maday, *Numerical analysis of nonlinear eigenvalue problems*, Journal of Scientific Computing **45** (2010), no. 1-3, 90–117.

[10] Eric Cances and Claude Le Bris, *On the convergence of SCF algorithms for the Hartree-Fock equations*, ESAIM: Mathematical Modelling and Numerical Analysis **34** (2000), no. 4, 749–774.

[11] Ziang Chen, Jianfeng Lu, Yulong Lu, and Xiangxiong Zhang, *On the Convergence of Sobolev Gradient Flow for the Gross–Pitaevskii Eigenvalue Problem*, SIAM Journal on Numerical Analysis **62** (2024), no. 2, 667–691.

[12] Ziang Chen, Jianfeng Lu, Yulong Lu, and Xiangxiong Zhang, *Fully discretized Sobolev gradient flow for the Gross–Pitaevskii eigenvalue problem*, Mathematics of Computation **94** (2025), no. 356, 2723–2760.

[13] Tianyang Chu, Xiaoying Dai, Jing Wu, and Aihui Zhou, *A gradient flow model for the Gross–Pitaevskii problem: mathematical and numerical analysis*, ESAIM: Mathematical Modelling and Numerical Analysis **60** (2026), no. 3, 1297–1326.

[14] Philippe G. Ciarlet, *The finite element method for elliptic problems*, Classics in Applied Mathematics, vol. 40, Society for Industrial and Applied Mathematics, Philadelphia, 2002.

[15] Ionut Danaila and Parimah Kazemi, *A new Sobolev gradient method for direct minimization of the Gross-Pitaevskii energy with rotation*, SIAM Journal on Scientific Computing **32** (2010), no. 5, 2447–2467.

[16] Ionut Danaila and Bartosz Protas, *Computation of ground states of the Gross-Pitaevskii functional via Riemannian optimization*, SIAM Journal on Scientific Computing **39** (2017), no. 6, B1102–B1129.

[17] Damek Davis and Wotao Yin, *A three-operator splitting scheme and its optimization applications*, Set-Valued and Variational Analysis **25** (2017), no. 4, 829–858.

[18] Zixu Feng, Lunxu Liu, and Qinglin Tang, *Energy dissipation and global convergence of a discrete normalized gradient flow for computing ground states of two-component Bose–Einstein condensates*, arXiv preprint arXiv:2510.19392 (2025).

[19] Zixu Feng and Qinglin Tang, *On preconditioned Riemannian gradient methods for minimizing the Gross–Pitaevskii energy functional: algorithms, global convergence and optimal local convergence rate*, arXiv preprint arXiv:2510.13516 (2025).

[20] Zixu Feng, Qinglin Tang, and Chushan Wang, *On the discrete normalized gradient flow for computing ground states of rotating Bose–Einstein condensates: energy dissipation and global convergence*, IMA Journal of Numerical Analysis (2025), doi:10.1093/imanum/draf100.

[21] Moritz Hauck, Yizhou Liang, and Daniel Peterseim, *Positivity preserving finite element method for the Gross–Pitaevskii ground state: discrete uniqueness and global convergence*, Numerische Mathematik **158** (2026), no. 3, 985–1014.

[22] Patrick Henning, *The dependency of spectral gaps on the convergence of the inverse iteration for a nonlinear eigenvector problem*, Mathematical Models and Methods in Applied Sciences **33** (2023), no. 07, 1517–1544.

[23] Patrick Henning and Elias Jarlebring, *The Gross–Pitaevskii equation and eigenvector nonlinearities: numerical methods and algorithms*, SIAM Review **67** (2025), no. 2, 256–317.

[24] Patrick Henning and Daniel Peterseim, *Sobolev gradient flow for the Gross-Pitaevskii eigenvalue problem: Global convergence and computational efficiency*, SIAM Journal on Numerical Analysis **58** (2020), no. 3, 1744–1772.

[25] Elias Jarlebring, Simen Kvaal, and Wim Michiels, *An inverse iteration method for eigenvalue problems with eigenvector nonlinearities*, SIAM Journal on Scientific Computing **36** (2014), no. 4, A1978–A2001.

[26] Parimah Kazemi and Michael Eckart, *Minimizing the Gross-Pitaevskii energy functional with the Sobolev gradient—Analytical and numerical results*, International Journal of Computational Methods **7** (2010), no. 03, 453–475.

[27] Hao Li and Xiangxiong Zhang, *On the monotonicity and discrete maximum principle of the finite difference implementation of C^0-Q^2 finite element method*, Numerische Mathematik **145** (2020), no. 2, 437–472.

[28] Elliott H. Lieb, Robert Seiringer, and Jakob Yngvason, *Bosons in a trap: a rigorous derivation of the Gross–Pitaevskii energy functional*, Physical Review A **61** (2000), no. 4, 043602.

[29] Xinyu Liu, Jie Shen, and Xiangxiong Zhang, *A simple GPU implementation of spectral-element methods for solving 3D Poisson type equations on rectangular domains and its applications*, Communications in Computational Physics **36** (2024), no. 5, 1157–1185.

[30] Xinyu Liu and Xiangxiong Zhang, *A Simple GPU-Accelerated Solver for the Schrödinger Operator with Applications to Ground States and Hamiltonian Simulation*, arXiv preprint arXiv:2605.20491 (2026).

[31] Lev P. Pitaevskii and Sandro Stringari, *Bose–Einstein Condensation*, International Series of Monographs on Physics, vol. 116, Oxford University Press, Oxford, 2003.

[32] Robert J Plemmons, *M-matrix characterizations. I—nonsingular M-matrices*, Linear Algebra and its Applications **18** (1977), no. 2, 175–188.

[33] Hugo Raguet, *A note on the forward-Douglas–Rachford splitting for monotone inclusion and convex optimization*, Optimization Letters **13** (2019), no. 4, 717–740.

[34] Parikshit Upadhyaya, Elias Jarlebring, and Emanuel H Rubensson, *A density matrix approach to the convergence of the self-consistent field iteration*, Numerical Algebra, Control and Optimization **11** (2021), no. 1, 99–115.

[35] Richard S. Varga, *Matrix iterative analysis*, second ed., Springer Series in Computational Mathematics, vol. 27, Springer, Berlin, 2000.

[36] Xinming Wu, Zaiwen Wen, and Weizhu Bao, *A regularized Newton method for computing ground states of Bose–Einstein condensates*, Journal of Scientific Computing **73** (2017), no. 1, 303–329.

[37] Jinchao Xu and Ludmil Zikatanov, *A monotone finite element scheme for convection-diffusion equations*, Mathematics of Computation **68** (1999), no. 228, 1429–1446.

[38] Chen Zhang, Patrick Henning, Mahima Yadav, and Wenbin Chen, *Convergence analysis of Sobolev gradient flows for the rotating Gross–Pitaevskii energy functional*, arXiv preprint arXiv:2510.15604 (2025).

[39] Ziyun Zhang, *Exponential convergence of Sobolev gradient descent for a class of nonlinear eigenproblems*, Communications in Mathematical Sciences **20** (2022), no. 2, 377–403.