

Riemannian Hessian analysis of Sobolev gradient flows for the Gross–Pitaevskii ground state problem

SHIXIN ZHENG

Department of Mathematics, University of Maryland, College Park, MD 20742, USA. Email: siuzheng@umd.edu

AND

XIANGXIONG ZHANG*

Department of Mathematics, Purdue University, 150 N. University Street, West Lafayette, IN 47907, USA

*Corresponding author: zhan1966@purdue.edu

Sobolev gradient flows for the Gross–Pitaevskii ground state problem are Riemannian gradient descent methods on the sphere under a metric induced by a discrete elliptic operator. We analyze their convergence through the condition number of the Riemannian Hessian at the ground state. Bounds of the Rayleigh quotient of the Riemannian Hessian give a uniform analysis of these methods for different metrics. The uniform analysis applies to the classical metrics, such as the shifted Laplacian and the energy-adaptive metrics. Inspired by this analysis, we also consider three new metrics. One of them is a Thomas–Fermi metric, whose condition number is bounded uniformly in the interaction strength for the one-dimensional harmonic trap and whose iteration count does not grow in the Thomas–Fermi regime in the numerical tests. The performance suggested by our analysis is consistent with three-dimensional numerical tests with up to one billion unknowns, for both Riemannian gradient descent and Riemannian conjugate gradient methods, whose practical performance is also related to the Hessian condition number. The Thomas–Fermi flow needs the fewest iterations, while the metrics inverted exactly by tensor-product solvers are the fastest in computational time on one GPU card.

Keywords: Gross–Pitaevskii ground state; Sobolev gradient flows; Riemannian optimization; Riemannian Hessian; preconditioning; Thomas–Fermi regime.

1. Introduction

1.1. The Gross–Pitaevskii ground state

The Gross–Pitaevskii ground state model (Gross, 1961; Pitaevskii, 1961) of a Bose–Einstein condensate of $\mathcal{N}$ identical bosons with scattering length a in a trapping potential V is given as a minimization over the whole space,

$$\mathcal{E}^{\text{GP}}(\phi) = \int_{\mathbb{R}^3} \left(|\nabla \phi|^2 + V |\phi|^2 + 4\pi a |\phi|^4 \right) d\mathbf{x} \quad \text{subject to} \quad \int_{\mathbb{R}^3} |\phi|^2 d\mathbf{x} = \mathcal{N},$$

and the infimum satisfies $E^{\text{GP}}(\mathcal{N}, a) = \mathcal{N} E^{\text{GP}}(1, \mathcal{N}a)$ (Lieb et al., 2000), so after rescaling to unit mass the interaction enters only through $\mathcal{N}a$. Numerical computations are usually carried out on a truncated bounded box $\Omega = (-L, L)^d$, which is justified by the exponential decay of the ground state (Bao and Cai, 2013, Theorem 2.5) and by the convergence of the truncated energy to the whole-space energy as $L \rightarrow \infty$ (Lieb et al., 2000, Lemma 2.2). Thus the size of the box is an approximation parameter

of the same kind as the mesh size. The rescaled problem is the constrained minimization

$$\min_{u \in H_0^1(\Omega), \int_{\Omega} |u|^2 dx = 1} \mathcal{E}(u), \quad \mathcal{E}(u) := \frac{1}{2} \int_{\Omega} (|\nabla u|^2 + V |u|^2) dx + \frac{\beta}{4} \int_{\Omega} |u|^4 dx, \quad (1)$$

on $\Omega = (-L, L)^d$, $d = 1, 2, 3$, with a trapping potential $V \geq 0$ and a coupling constant $\beta > 0$, which is the defocusing case. Its Euler–Lagrange equation is the nonlinear eigenvalue problem: find $\lambda \in \mathbb{R}$ and $u \in H_0^1(\Omega)$ with $\|u\|_{L^2} = 1$ such that

$$(\nabla u, \nabla v) + (Vu, v) + \beta(|u|^2 u, v) = \lambda(u, v), \quad \forall v \in H_0^1(\Omega), \quad (2)$$

where $(\cdot, \cdot)$ is the L^2 inner product on Ω . For $\beta > 0$ the minimizer exists, is unique up to sign and can be chosen strictly positive, and it is the eigenfunction of this eigenvalue problem for the smallest eigenvalue λ , the chemical potential (Cancès et al., 2010; Henning and Jarlebring, 2025). Taking $v = u$ in (2) shows that the eigenvalue of a normalized eigenfunction u is determined by u through

$$\lambda(u) := 2\mathcal{E}(u) + \frac{\beta}{2} \int_{\Omega} |u|^4 dx. \quad (3)$$

The right-hand side is defined for every $u \in H_0^1(\Omega)$ with $\|u\|_{L^2} = 1$. The coupling constant $\beta = 8\pi a_s \mathcal{N}$ (Lieb et al., 2000, (2.4)) is proportional to the number of particles, and typical experiments with $a \approx 10^{-3}$ and $\mathcal{N}$ between 10^2 and 10^7 correspond to β between about 2.5 and 2.5×10^5 (Bao and Cai, 2013, Section 1.3). In the regime of large β the ground state approaches the Thomas–Fermi profile $\sqrt{(\lambda - V)_+ / \beta}$, where $(t)_+ := \max\{t, 0\}$, with a boundary layer at the edge of its support that the mesh must resolve.

The fully discretized problem thus has three parameters of interest, the mesh size h , the box size L and the coupling constant β . The limits $h \rightarrow 0$ and $L \rightarrow \infty$ are needed for the discrete ground state to approximate the ground state on $\mathbb{R}^d$, and large β is the physical regime. Computation is more difficult for smaller h , larger L and larger β .

1.2. Sobolev gradient flows as Riemannian gradient descent

The Fréchet derivative $\mathcal{E}'(v) \in H^{-1}(\Omega)$, $\langle \mathcal{E}'(v), w \rangle = (\nabla v, \nabla w) + (Vv, w) + \beta(|v|^2 v, w)$, must be represented in a Hilbert space before it can serve as a descent direction. To do so, we fix an inner product $(\cdot, \cdot)_X$ on $H_0^1(\Omega)$. Classical examples are the H^1 inner product and the energy-adaptive inner product (Henning and Peterseim, 2020),

$$(v, w)_{H^1} := (\nabla v, \nabla w) + (v, w), \quad (v, w)_{a_u} := (\nabla v, \nabla w) + (Vv, w) + \beta(u^2 v, w), \quad (4)$$

where $(\cdot, \cdot)_{a_u}$ depends on the current iterate u . Every inner product used in this paper can be computed via an elliptic operator A_X , $(v, w)_X = (A_X v, w)$, for instance $A_{a_u} = -\Delta + V + \beta u^2$ for the second inner product in (4). The Sobolev gradient $\nabla_X \mathcal{E}(v)$ is the Riesz representative of $\mathcal{E}'(v)$ with respect to $(\cdot, \cdot)_X$ (Neuberger, 1997), defined by $(\nabla_X \mathcal{E}(v), w)_X = \langle \mathcal{E}'(v), w \rangle$ for all $w \in H_0^1(\Omega)$. Thus the Sobolev gradient is one elliptic solve, $\nabla_X \mathcal{E}(v) = A_X^{-1} \mathcal{E}'(v)$. For the a_u inner product, $(u, w)_{a_u} = \langle \mathcal{E}'(u), w \rangle$ for all w by (4), so $\nabla_{a_u} \mathcal{E}(u) = u$ and no solve is needed.

Projecting the Sobolev gradient onto the tangent space $\{w \in H_0^1(\Omega) : (u, w) = 0\}$ with respect to $(\cdot, \cdot)_X$ and renormalizing gives the Sobolev gradient descent

$$u^{(k+1)} = \frac{u^{(k)} - \tau_k P_{u^{(k)}}^X \nabla_X \mathcal{E}(u^{(k)})}{\|u^{(k)} - \tau_k P_{u^{(k)}}^X \nabla_X \mathcal{E}(u^{(k)})\|_{L^2(\Omega)}}, \quad (5)$$

with the X -orthogonal projection P_u^X onto the tangent space and a step size $\tau_k > 0$. The projection is one more elliptic solve. In fact, $A_X^{-1}u$ spans the X -orthogonal complement of the tangent space, since $(A_X^{-1}u, w)_X = (u, w) = 0$ for every w in it, and thus

$$P_u^X v = v - \frac{(u, v)}{(u, A_X^{-1}u)} A_X^{-1}u, \quad P_u^X \nabla_X \mathcal{E}(u) = A_X^{-1} \mathcal{E}'(u) - \frac{(u, A_X^{-1} \mathcal{E}'(u))}{(u, A_X^{-1}u)} A_X^{-1}u. \quad (6)$$

For the a_u metric, $\nabla_{a_u} \mathcal{E}(u) = u$ and the single solve $A_{a_u}^{-1}u$ suffices. For other metrics, one step of (5) thus costs two elliptic solves with A_X , one for the gradient and one for the projection. Table 1 collects three classical metrics, the modified H^1 , a_0 and a_u metrics, and three new metrics which we consider in this paper. The new metrics are variants of the classical ones and are inspired by the Riemannian Hessian analysis in Section 3.

classical metric	elliptic operator A_X	solves/iter.	discussed in
modified H^1	$-\Delta + \alpha$	2	Chen et al. (2025)
a_0	$-\Delta + V$	2	Danaila and Kazemi (2010) ; Danaila and Protas (2017) ; Chen et al. (2024)
a_u	$-\Delta + V + \beta u^2$	1	Henning and Peterseim (2020) ; Hauck et al. (2026)
new metric	elliptic operator A_X	solves/iter.	discussed in
modified a_0	$-\Delta + V_1$	2	Section 3.3
modified a_u	$-\Delta + V + 3\beta u^2$	2	Section 3.4
Thomas–Fermi	$-\Delta + 2\beta u^2 + (V - \lambda(u))_+ + \ell \lambda(u)^{1/3}$	2	Section 3.5

TABLE 1 Here $\alpha \geq 0$ is a constant, $V_1(\mathbf{x}) = \sum_{j=1}^d v_j(x_j)$ is a separable part of the potential, $\lambda(u)$ is defined by (3) for the current iterate u and $\ell > 0$ is a fixed parameter.

The iteration (5) is a Riemannian gradient descent on the sphere, in which the projected Sobolev gradient is the Riemannian gradient for the metric induced by $(\cdot, \cdot)_X$ ([Danaila and Protas, 2017](#)) and the normalization is a retraction ([Absil et al., 2008](#), Sect. 4.1). It can also be regarded as a preconditioned projected gradient descent with the preconditioner A_X^{-1} ([Antoine et al., 2017](#); [Henning and Jarlebring, 2025](#)). For the a_u metric, the step with $\tau_k = 1$ is an inverse iteration for the linearized operator, and $\tau_k \in (0, 1)$ gives the damped inverse iteration ([Henning and Peterseim, 2020](#); [Henning and Jarlebring, 2025](#)). After discretization in space, with quadrature for the nonlinear terms, the metrics become matrices, and Section 2 formulates the iteration as a Riemannian gradient descent on a sphere in $\mathbb{R}^N$.

1.3. Main results and related work

In this paper we analyze the Riemannian Hessian of the discrete energy at the ground state for an arbitrary metric induced by a discrete elliptic operator, and we compare metrics through its conditioning, as in our earlier work on fixed-rank matrix optimization ([Zheng et al., 2025](#)). At a critical point this

Hessian is the second variation of the discrete energy along the normalization retraction (Lemma 3.1). It is thus the same quadratic form for every metric, and its Rayleigh quotient is the ratio of a metric-independent constrained Hessian to the metric operator. This gives two-sided bounds at the ground state (Theorem 3.4), closed-form bounds on the condition number κ (Corollary 3.5) and local rates, in particular the rate $(\kappa - 1)/(\kappa + 1)$ of the fixed-step iteration at the optimal step (Theorem 4.2 and Proposition 4.3). For the a_u metric, the lower bound on the spectrum is asymptotically exact in the Thomas–Fermi regime for the one-dimensional harmonic trap, where the condition number grows like the two-thirds power of the ground state eigenvalue (Corollary 3.11). The corresponding bound based on the spectral gap alone is not sharp (Henning, 2023, Remark 3.3), and Appendix A compares both bounds with exact spectra in one, two and three dimensions. For the normalized a_0 flow, an explicit bound for the local contraction rate is listed as open in Henning and Jarlebring (2025, Section 5.2.2), and we give one for the fully discretized flow (Corollary 3.5 and Theorem 4.2). The bound for the modified H^1 metric grows with the box size L , as do its iteration counts in the tests. We also prove convergence of the whole iterate sequence to a critical point from any initial value (Theorem 4.4).

A metric whose operator is closer to the constrained Hessian is better conditioned. Inspired by this analysis, we also consider three new metrics. The modified a_0 metric keeps only the separable part V_1 of the potential, and its operator $-\Delta + V_1$ is inverted exactly by tensor-product solvers. It is motivated by Liu and Zhang (2026), where $(-\Delta + V_1)^{-1}$ preconditions the inner solves of the a_u flow and gives mesh- and domain-independent conditioning when V_1 is confining. The Riemannian Hessian analysis shows that the metric inherits the robustness in L . The modified a_u metric asymptotically halves the condition number of the a_u metric (Proposition 3.8). The Thomas–Fermi metric is a Thomas–Fermi approximation of the constrained Hessian. For the one-dimensional harmonic trap, it is not uniformly equivalent to the a_u metric, and its condition number is bounded uniformly in β (Proposition 3.10 and Remark 3.12). In three-dimensional tests with up to 10^9 unknowns (Section 5), the Thomas–Fermi flow needs 8 iterations against 35 for the a_u flow, and its iteration count does not grow over a factor of 64 in β . Nevertheless, the modified H^1 and modified a_0 metrics are the fastest in wall clock time, because their operators are separable. The modified a_0 metric remains the fastest as L grows.

The literature on computing Gross–Pitaevskii ground states is extensive, and we refer to the review (Henning and Jarlebring, 2025). The gradient flow with discrete normalization (Bao and Du, 2004), a semi-implicit discretization of the projected L^2 gradient flow, remains the most widely used method. The classical metrics of Table 1 are those of the H^1 flow (Kazemi and Eckart, 2010; Chen et al., 2025), of the a_0 flow and its Riemannian gradient and conjugate gradient versions (Danaila and Kazemi, 2010; Danaila and Protas, 2017), and of the a_u flow (Henning and Peterseim, 2020), which was extended to Kohn–Sham energies in Altmann et al. (2022). Other approaches include preconditioned gradient methods in the L^2 metric (Antoine et al., 2017), second-order flows (Chen et al., 2023), the J -method (Altmann et al., 2021) and Riemannian Newton methods (Altmann et al., 2024), and their convergence theory has been developed mostly at the continuous level (Kazemi and Eckart, 2010; Faou and Jézéquel, 2018; Henning and Peterseim, 2020; Zhang, 2022; Chen et al., 2024; Henning and Yadav, 2025a; Zhang et al., 2025). The basic inverse iteration is the a_u flow with the step $\tau_k = 1$. It converges locally at the rate λ_1/λ_2 , the ratio of the first two eigenvalues of the linearized operator (Henning, 2023, Theorem 3.2). Its sharp rate is the largest modulus $|\mu_1|$ of the eigenvalues of a weighted eigenvalue problem, and $|\mu_1| < \lambda_1/\lambda_2$ for every $\beta > 0$ (Henning, 2023, Remark 3.3). Up to a shift, this weighted eigenvalue problem is the eigenvalue problem of the Riemannian Hessian under the a_u metric (Section 4.1). For the damped inverse iteration with a fixed damping parameter $\tau \in (0, 1]$ the rate is the convex combination $|1 - \tau| + \tau|\mu_1| \leq |1 - \tau| + \tau\lambda_1/\lambda_2$ (Henning, 2023, Theorem 6.2), (Henning and Jarlebring, 2025, Theorem 5.17), and Theorem 4.2 gives it as $\max_j |1 - \tau + \tau\mu_j|$, the

sharper value noted in [Henning \(2023, Section 6.1\)](#) (Section 4.1). For the gradient flow with discrete normalization, energy dissipation and global convergence were established recently ([Feng et al., 2025](#)), and a unified framework of preconditioned Riemannian gradient methods with a local rate derived from a Polyak–Łojasiewicz inequality was proposed in [Feng and Tang \(2025\)](#). Riemannian metrics based on regularized Lagrangian Hessians have also been studied ([Altmann et al., 2026](#); [Feng and Tang, 2025](#)). Results for the fully discretized problem are fewer: a priori error estimates ([Cancès et al., 2010](#); [Henning et al., 2014](#); [Henning and Persson, 2023](#); [Henning and Yadav, 2025b](#)), global convergence and a mesh-independent local rate of the discrete modified H^1 flow ([Chen et al., 2025](#)), and positivity and uniqueness of the discrete ground state with global convergence of the discrete a_u flow from nonnegative data ([Hauck et al., 2026](#)). For monotone discrete Laplacians, our recent work ([Li et al., 2026](#)) proves global convergence to the positive discrete ground state from every positive normalized initial vector for a Davis–Yin operator splitting scheme, which treats the potential and the interaction explicitly and inverts only a shifted Laplacian in each iteration. The Thomas–Fermi asymptotics used in Corollary 3.11 were established in [Gallo and Pelinovsky \(2011\)](#) for radial harmonic traps, and the boundary layer analysis was extended to general potentials in two dimensions in [Karali and Sourdís \(2015\)](#).

1.4. Organization of the paper

Section 2 sets up the discrete problem, the six metrics and the iteration. Section 3 contains the analysis of the Hessian at a critical point, the two-sided bounds, the three new metrics and the Thomas–Fermi asymptotics of the a_u metric. Section 4 turns the bounds into local rates and records global convergence, Section 5 reports the three-dimensional experiments and Section 6 concludes. Appendix A tests the bounds against exact spectra of the Riemannian Hessian, and Appendix B contains the Riemannian connection, the Hessian at an arbitrary point and the vector transport.

2. Riemannian gradient descent

2.1. Discretization

We discretize (1) by the Q^k finite element method with tensor-product Gauss–Lobatto quadrature on a rectangular mesh, following [Chen et al. \(2025, Section 2\)](#), to which we refer for the details and for the P^k element on a simplicial mesh. Let $\{\mathbf{x}_i\}_{i=1}^N$ be the interior quadrature nodes, which are also the interpolation nodes of the Lagrange basis $\{\phi_i\}_{i=1}^N$ of the finite element space $V_h^0 \subset H_0^1(\Omega)$, and let $w_i > 0$ be the quadrature weights. Every $u_h \in V_h^0$ is identified with its vector of nodal values $\mathbf{u} = [u_h(\mathbf{x}_1), \dots, u_h(\mathbf{x}_N)]^T \in \mathbb{R}^N$, and the discrete L^2 inner product is

$$\langle \mathbf{u}, \mathbf{v} \rangle_h := \mathbf{u}^T \mathbb{M} \mathbf{v} = \sum_{i=1}^N w_i u_i v_i, \quad \|\mathbf{u}\|_h := \langle \mathbf{u}, \mathbf{u} \rangle_h^{1/2}, \quad \mathbb{M} := \text{diag}(w_1, \dots, w_N).$$

With the potential matrix $\mathbb{V} := \text{diag}(V(\mathbf{x}_1), \dots, V(\mathbf{x}_N))$ and the stiffness matrix $\mathbb{S}_{ij} := \langle \nabla \phi_i, \nabla \phi_j \rangle_h$, the energy (1) with all integrals replaced by quadrature is the fully discrete energy on $\mathbb{R}^N$,

$$E(\mathbf{u}) = \frac{1}{2} \mathbf{u}^T \mathbb{S} \mathbf{u} + \frac{1}{2} \mathbf{u}^T \mathbb{M} \mathbb{V} \mathbf{u} + \frac{\beta}{4} (\mathbf{u}^2)^T \mathbb{M} \mathbf{u}^2, \quad (7)$$

to be minimized over the sphere $\mathcal{M} := \{\mathbf{u} \in \mathbb{R}^N : \langle \mathbf{u}, \mathbf{u} \rangle_h = 1\}$, where powers of vectors such as $\mathbf{u}^2$ are taken componentwise. Its Euler–Lagrange equation is the discrete eigenproblem

$$A_\beta(\mathbf{u})\mathbf{u} = \lambda \mathbf{u}, \quad A_\beta(\mathbf{u}) := -\Delta_h + \mathbb{V} + \beta \text{diag}(\mathbf{u}^2), \quad \langle \mathbf{u}, \mathbf{u} \rangle_h = 1, \quad (8)$$

that is, $\mathbb{S}\mathbf{u} + \mathbb{M}\mathbb{V}\mathbf{u} + \beta\mathbb{M}\mathbf{u}^3 = \lambda\mathbb{M}\mathbf{u}$, where $-\Delta_h := \mathbb{M}^{-1}\mathbb{S}$. The matrix $-\Delta_h$ is not symmetric in general. However, every operator $-\Delta_h + \mathbb{D}$ with a diagonal $\mathbb{D}$, including $-\Delta_h$ itself, is self-adjoint with respect to $\langle \cdot, \cdot \rangle_h$, since $\mathbb{M}(-\Delta_h + \mathbb{D}) = \mathbb{S} + \mathbb{M}\mathbb{D}$ is symmetric. Such an operator A has real eigenvalues $\lambda_1(A) \leq \dots \leq \lambda_N(A)$, which are the eigenvalues of the symmetric pencil $(\mathbb{M}A, \mathbb{M})$, and the Courant–Fischer min-max principle holds for them in $\langle \cdot, \cdot \rangle_h$. We also write $\mathbb{I}$ for the identity matrix and $A_{3\beta}(\mathbf{u}) := -\Delta_h + \mathbb{V} + 3\beta\text{diag}(\mathbf{u}^2)$ for the derivative of $\mathbf{u} \mapsto A_\beta(\mathbf{u})\mathbf{u}$. For convenience, we use A_β and $A_{3\beta}$ to denote $A_\beta(\mathbf{u}^*)$ and $A_{3\beta}(\mathbf{u}^*)$, where $\mathbf{u}^*$ is the ground state.

On a uniform mesh, the Q^1 scheme with this quadrature reduces to the second-order finite difference scheme in any dimension, and so does the P^1 scheme in one dimension. Part of the analysis uses the sign structure of the scheme.

ASSUMPTION 2.1 (Monotone scheme) The stiffness matrix $\mathbb{S}$ is an irreducible Stieltjes matrix, that is, symmetric positive definite with nonpositive off-diagonal entries and an irreducible sparsity pattern, and its row sums are nonnegative.

Assumption 2.1 holds for the second-order finite difference scheme in any dimension and for the P^1 finite element method with mass lumping on a simplicial mesh satisfying an angle condition, which in two dimensions means a Delaunay triangulation (Chen et al., 2025). The Perron–Frobenius theorem and the convexity of the energy in $\mathbf{u}^2$ give the following lemma (Chen et al., 2025, Section 3).

LEMMA 2.2 (Simple smallest eigenvalue and positive ground state) Under Assumption 2.1, for every $\mathbf{u} \in \mathbb{R}^N$ the matrices $\mathbb{M}A_\beta(\mathbf{u}) = \mathbb{S} + \mathbb{M}(\mathbb{V} + \beta\text{diag}(\mathbf{u}^2))$ and $\mathbb{M}A_{3\beta}(\mathbf{u})$ are irreducible Stieltjes matrices. Thus $A_\beta(\mathbf{u})$ and $A_{3\beta}(\mathbf{u})$ are nonsingular M-matrices with entrywise positive inverses, and their smallest eigenvalues are simple with entrywise positive eigenvectors. In particular, the discrete ground state $\mathbf{u}^*$ is unique up to sign and strictly positive, and it is an eigenvector of the simple smallest eigenvalue $\lambda_1(A_\beta)$ of A_β .

2.2. Riemannian metrics

A Riemannian metric on $\mathbb{R}^N$ is specified at each point $\mathbf{u} \in \mathbb{R}^N$ by a matrix $B(\mathbf{u})$, self-adjoint and positive definite with respect to $\langle \cdot, \cdot \rangle_h$, through

$$g_{\mathbf{u}}(\xi, \eta) := \langle B(\mathbf{u})\xi, \eta \rangle_h = \xi^T \mathbb{M}B(\mathbf{u})\eta, \quad \xi, \eta \in \mathbb{R}^N. \quad (9)$$

We call $B(\mathbf{u})$ the metric operator. The metric is called *Euclidean*, or constant, if B does not depend on $\mathbf{u}$, and *$\mathbf{u}$ -dependent* otherwise. The discrete counterparts of the six metrics of Table 1, with $-\Delta_h$ in place of $-\Delta$, are

$$\begin{aligned} \text{modified } H^1: B &= -\Delta_h + \alpha\mathbb{I}, \quad \alpha \geq 0, & \text{modified } a_0: B &= A_1 := -\Delta_h + \mathbb{V}_1, \\ a_0: B &= -\Delta_h + \mathbb{V}, & \text{modified } a_u: B(\mathbf{u}) &= A_{3\beta}(\mathbf{u}), \\ a_u: B(\mathbf{u}) &= A_\beta(\mathbf{u}), & \text{Thomas–Fermi: } B(\mathbf{u}) &= B_{\text{TF}}(\mathbf{u}), \end{aligned} \quad (10)$$

where

$$B_{\text{TF}}(\mathbf{u}) := -\Delta_h + 2\beta\text{diag}(\mathbf{u}^2) + \text{diag}((\mathbb{V} - \lambda(\mathbf{u}))_+) + \ell\lambda(\mathbf{u})^{1/3}\mathbb{I} \quad (11)$$

with the positive part $(\cdot)_+$ taken entrywise, a parameter $\ell > 0$ and the discrete counterpart of (3),

$$\lambda(\mathbf{u}) := \langle A_\beta(\mathbf{u})\mathbf{u}, \mathbf{u} \rangle_h = 2E(\mathbf{u}) + \frac{\beta}{2} \langle \mathbf{u}^2, \mathbf{u}^2 \rangle_h, \quad (12)$$

which is nonnegative for every $\mathbf{u} \in \mathbb{R}^N$. For the modified a_0 metric, we split $\mathbb{V} = \mathbb{V}_1 + \mathbb{V}_2$, where $\mathbb{V}_1 \geq 0$ samples a *separable* potential $V_1(\mathbf{x}) = \sum_{j=1}^d v_j(x_j)$ and $\mathbb{V}_2$ is bounded uniformly in h . Each of the six operators has the form $B(\mathbf{u}) = -\Delta_h + \mathbb{D}$ with a nonnegative diagonal matrix $\mathbb{D}$. Thus $\mathbb{M}B(\mathbf{u}) = \mathbb{S} + \mathbb{M}\mathbb{D}$ is symmetric positive definite, and (9) is a metric on all of $\mathbb{R}^N$. Applying $B(\mathbf{u})^{-1}$ requires one sparse linear solve with $\mathbb{S} + \mathbb{M}\mathbb{D}$. On rectangular meshes with Q^k elements and a separable diagonal $\mathbb{D}$, as in the modified H^1 and modified a_0 metrics, this solve is carried out exactly by tensor-product direct solvers at $O(N^{1+1/d})$ cost (Liu et al., 2024; Liu and Zhang, 2026), while the other four operators need an inner iterative solver.

For a map F on $\mathbb{R}^N$, we write $DF(\mathbf{x})[\xi] := \frac{d}{dt}F(\mathbf{x} + t\xi)|_{t=0}$ for its directional derivative. The gradient of E in $(\mathbb{R}^N, g)$ (Absil et al., 2008, (3.31)) is the Riesz representative of $DE(\mathbf{u})[\eta] = \langle A_\beta(\mathbf{u})\mathbf{u}, \eta \rangle_h$ in $g_{\mathbf{u}}$, which is the discrete Sobolev gradient

$$\nabla_X E(\mathbf{u}) = B(\mathbf{u})^{-1} A_\beta(\mathbf{u})\mathbf{u}. \quad (13)$$

For the a_u metric, $\nabla_X E(\mathbf{u}) = \mathbf{u}$.

2.3. The manifold and the iteration

The sphere $\mathcal{M}$ is an embedded submanifold of $(\mathbb{R}^N, g)$ by the submersion theorem (Absil et al., 2008, Prop. 3.3.3), with tangent and normal spaces

$$T_{\mathbf{u}}\mathcal{M} = \{\xi \in \mathbb{R}^N : \langle \xi, \mathbf{u} \rangle_h = 0\}, \quad N_{\mathbf{u}}\mathcal{M} = \{sB(\mathbf{u})^{-1}\mathbf{u} : s \in \mathbb{R}\}. \quad (14)$$

The tangent space is the kernel of the derivative $\xi \mapsto 2\langle \xi, \mathbf{u} \rangle_h$ of the constraint $\langle \mathbf{u}, \mathbf{u} \rangle_h = 1$ (Absil et al., 2008, (3.19)). The normal space is the $g_{\mathbf{u}}$ -orthogonal complement of the tangent space (Absil et al., 2008, Sect. 3.6.1). It is spanned by $B(\mathbf{u})^{-1}\mathbf{u}$, since $g_{\mathbf{u}}(B(\mathbf{u})^{-1}\mathbf{u}, \xi) = \langle \mathbf{u}, \xi \rangle_h = 0$ for every $\xi \in T_{\mathbf{u}}\mathcal{M}$. The $g_{\mathbf{u}}$ -orthogonal projection onto $T_{\mathbf{u}}\mathcal{M}$ and the Riemannian gradient of E on $\mathcal{M}$ are thus

$$\begin{aligned} P_{\mathbf{u}}\mathbf{w} &= \mathbf{w} - \frac{\langle \mathbf{u}, \mathbf{w} \rangle_h}{\langle \mathbf{u}, B(\mathbf{u})^{-1}\mathbf{u} \rangle_h} B(\mathbf{u})^{-1}\mathbf{u}, \\ \text{grad} E(\mathbf{u}) &= P_{\mathbf{u}}\nabla_X E(\mathbf{u}) = B(\mathbf{u})^{-1}(A_\beta(\mathbf{u})\mathbf{u} - \lambda_B(\mathbf{u})\mathbf{u}), \end{aligned} \quad (15)$$

where $\lambda_B(\mathbf{u}) := \langle \mathbf{u}, B(\mathbf{u})^{-1}A_\beta(\mathbf{u})\mathbf{u} \rangle_h / \langle \mathbf{u}, B(\mathbf{u})^{-1}\mathbf{u} \rangle_h$ is the metric-weighted Rayleigh quotient. The Riemannian gradient is the projection of the ambient gradient (13) onto the tangent space (Absil et al., 2008, (3.37)), and both formulas are the matrix form of (6). Thus $\text{grad} E(\mathbf{u}) = 0$ exactly when $\mathbf{u}$ is an eigenvector of $A_\beta(\mathbf{u})$, that is, a solution of (8), and pairing with $\mathbf{u}$ identifies the eigenvalue as $\lambda(\mathbf{u})$ of (12), where also $\lambda_B(\mathbf{u}) = \lambda(\mathbf{u})$. The Riemannian gradient descent for minimizing E over $\mathcal{M}$ with step size $\tau_k > 0$ is

$$\mathbf{u}_{k+1} = R_{\mathbf{u}_k}(-\tau_k \text{grad} E(\mathbf{u}_k)), \quad R_{\mathbf{u}}(\xi) := \frac{\mathbf{u} + \xi}{\|\mathbf{u} + \xi\|_h} = \frac{\mathbf{u} + \xi}{\sqrt{1 + \|\xi\|_h^2}}, \quad \xi \in T_{\mathbf{u}}\mathcal{M}, \quad (16)$$

where the normalization $R_{\mathbf{u}}$ is a retraction (Absil et al., 2008, Def. 4.1.1 and Example 4.1.1), that is, $R_{\mathbf{u}}(0) = \mathbf{u}$ and $\frac{d}{dt}R_{\mathbf{u}}(t\xi)|_{t=0} = \xi$ for every $\xi \in T_{\mathbf{u}}\mathcal{M}$. The retraction replaces the exponential map of $(\mathcal{M}, g)$ (Absil et al., 2008, Sect. 5.4), which has no closed form for the $\mathbf{u}$ -dependent metrics. This iteration is the matrix form of (5). The retraction $R_{\mathbf{u}}$ maps $T_{\mathbf{u}}\mathcal{M}$ bijectively onto the open hemisphere $\{\mathbf{w} \in \mathcal{M} : \langle \mathbf{w}, \mathbf{u} \rangle_h > 0\}$, with inverse $\mathbf{w} \mapsto \mathbf{w} / \langle \mathbf{w}, \mathbf{u} \rangle_h - \mathbf{u}$. In fact, $\langle R_{\mathbf{u}}(\xi), \mathbf{u} \rangle_h = (1 + \|\xi\|_h^2)^{-1/2} > 0$, and pairing $R_{\mathbf{u}}(\xi) = \mathbf{w}$ with $\mathbf{u}$ gives $\mathbf{u} + \xi = \mathbf{w} / \langle \mathbf{w}, \mathbf{u} \rangle_h$.

3. Riemannian Hessian analysis

At a critical point, the Rayleigh quotient of the Riemannian Hessian in a tangent direction η is the second variation of E in this direction divided by $g_{\mathbf{u}}(\eta, \eta)$. Only the denominator depends on the metric (Section 3.1). Comparing the numerator with the denominator gives two-sided bounds at the ground state for all metrics at once (Section 3.2), following the approach in Zheng et al. (2025). Inspired by this comparison, we also consider three new metrics (Sections 3.3 to 3.5). Section 3.5 also derives the sharp Thomas–Fermi asymptotics of the a_u metric from Gallo and Pelinovsky (2011). Appendix A tests the bounds against exact spectra of the Riemannian Hessian.

3.1. The Hessian at a critical point

The Riemannian Hessian of E on $\mathcal{M}$ at $\mathbf{u}$ is the linear map $\text{Hess}E(\mathbf{u})$ of $T_{\mathbf{u}}\mathcal{M}$ into itself defined by $\text{Hess}E(\mathbf{u})[\xi] = \nabla_{\xi} \text{grad}E$, where ∇ is the Riemannian connection of $(\mathcal{M}, g)$ (Absil et al., 2008, Def. 5.5.1), and its Rayleigh quotient is

$$\rho_{\mathbf{u}}(\eta) := \frac{g_{\mathbf{u}}(\text{Hess}E(\mathbf{u})[\eta], \eta)}{g_{\mathbf{u}}(\eta, \eta)}, \quad \eta \in T_{\mathbf{u}}\mathcal{M} \setminus \{0\}. \quad (17)$$

The Hessian is self-adjoint with respect to $g_{\mathbf{u}}$ (Absil et al., 2008, Prop. 5.5.3), so its eigenvalues are real and its extreme eigenvalues are the extreme values of $\rho_{\mathbf{u}}$.

LEMMA 3.1 (Second variation at a critical point) Let $\mathbf{u}^* \in \mathcal{M}$ be a critical point of E on $\mathcal{M}$, so that $A_{\beta}(\mathbf{u}^*)\mathbf{u}^* = \lambda(\mathbf{u}^*)\mathbf{u}^*$, and let $\eta \in T_{\mathbf{u}^*}\mathcal{M}$. Then, for every metric (9) that is smooth near $\mathbf{u}^*$ and every smooth curve $\mathbf{u}(t)$ on $\mathcal{M}$ with $\mathbf{u}(0) = \mathbf{u}^*$ and $\mathbf{u}'(0) = \eta$,

$$g_{\mathbf{u}^*}(\text{Hess}E(\mathbf{u}^*)[\eta], \eta) = \left. \frac{d^2}{dt^2} E(\mathbf{u}(t)) \right|_{t=0} = \langle (A_{3\beta}(\mathbf{u}^*) - \lambda(\mathbf{u}^*)\mathbb{I})\eta, \eta \rangle_h. \quad (18)$$

Proof The chain rule and the compatibility of ∇ with the metric (Absil et al., 2008, Theorem 5.3.1) give $\frac{d^2}{dt^2} E(\mathbf{u}(t)) = g_{\mathbf{u}(t)}(\text{Hess}E(\mathbf{u}(t))[\mathbf{u}'(t)], \mathbf{u}'(t)) + g_{\mathbf{u}(t)}(\text{grad}E(\mathbf{u}(t)), \frac{D}{dt}\mathbf{u}'(t))$, where $\frac{D}{dt}$ is the covariant derivative along the curve (Absil et al., 2008, Sect. 5.4). Along the curve, these two properties are the chain rule and the product rule of Boumal (2023, Theorem 5.29). The second term vanishes at $t = 0$ because $\text{grad}E(\mathbf{u}^*) = 0$. This is the first equality. For curves $t \mapsto R_{\mathbf{u}^*}(t\eta)$ generated by a retraction, it is also a consequence of Absil et al. (2008, Prop. 5.5.6). For the second equality, E is a polynomial on $\mathbb{R}^N$, so $\frac{d^2}{dt^2} E(\mathbf{u}(t))|_{t=0} = D^2E(\mathbf{u}^*)[\eta, \eta] + DE(\mathbf{u}^*)[\mathbf{u}''(0)]$ with $D^2E(\mathbf{u}^*)[\eta, \eta] = \langle A_{3\beta}(\mathbf{u}^*)\eta, \eta \rangle_h$ and $DE(\mathbf{u}^*)[\mathbf{w}] = \langle A_{\beta}(\mathbf{u}^*)\mathbf{u}^*, \mathbf{w} \rangle_h = \lambda(\mathbf{u}^*)\langle \mathbf{u}^*, \mathbf{w} \rangle_h$ by the eigenvalue equation, while differentiating $\langle \mathbf{u}(t), \mathbf{u}(t) \rangle_h = 1$ twice at $t = 0$ gives $\langle \mathbf{u}^*, \mathbf{u}''(0) \rangle_h = -\langle \eta, \eta \rangle_h$. $\square$

The right-hand side of (18) depends neither on the metric nor on the retraction. Only the denominator of (17) depends on the metric, and at the ground state $\mathbf{u}^*$, where $\lambda(\mathbf{u}^*) = \lambda_1(A_{\beta})$ under the assumption of Section 3.2,

$$\rho_{\mathbf{u}^*}(\eta) = \frac{\langle H^*\eta, \eta \rangle_h}{\langle B^*\eta, \eta \rangle_h}, \quad H^* := A_{3\beta}(\mathbf{u}^*) - \lambda_1(A_{\beta}(\mathbf{u}^*))\mathbb{I}, \quad B^* := B(\mathbf{u}^*), \quad (19)$$

where the constrained Hessian H^* is independent of the metric. Bounding the Rayleigh quotient thus amounts to comparing the metric operator B^* with H^* . Appendix B gives the Hessian at arbitrary points, which requires the Riemannian connection, and its Lemma B.2 reduces to (18) at a critical point.

Lemma 3.1 requires a metric that is smooth near $\mathbf{u}^*$. This holds for the first five metrics of (10), and for the Thomas–Fermi metric exactly when $\mathbb{V}_i \neq \lambda_1(A_\beta)$ for every node i (Appendix B). If $\mathbb{V}_i = \lambda_1(A_\beta)$ at some node, then B_{TF} is Lipschitz but not differentiable at $\mathbf{u}^*$, and we define $\text{Hess } E(\mathbf{u}^*)$ as the operator on $T_{\mathbf{u}^*}\mathcal{M}$ that is self-adjoint with respect to $g_{\mathbf{u}^*}$ and satisfies $g_{\mathbf{u}^*}(\text{Hess } E(\mathbf{u}^*)[\eta], \xi) = \langle H^*\eta, \xi \rangle_h$ for all $\eta, \xi \in T_{\mathbf{u}^*}\mathcal{M}$. Whenever the metric is smooth near $\mathbf{u}^*$, this operator is the Riemannian Hessian by Lemma 3.1 and the polarization identity (Absil et al., 2008, (5.36)). Thus (18) and (19) hold for every metric, and all results of this section use only (19).

3.2. Two-sided bounds at the ground state

Throughout this section $\mathbf{u}^*$ is the ground state. We assume that $\mathbf{u}^*$ is an eigenvector of the smallest eigenvalue $\lambda_1(A_\beta)$ of A_β and that this eigenvalue is simple, so that $\lambda(\mathbf{u}^*) = \lambda_1(A_\beta)$. Both properties hold under Assumption 2.1 by Lemma 2.2. We write

$$Z := H^* - B^*, \quad \zeta_+ := \max_i (Z_{ii})_+, \quad \zeta_- := \max_i (-Z_{ii})_+, \quad (20)$$

for the difference between the two operators and its extreme entries. It is a diagonal matrix for all six metrics and is listed in Table 2. We also write

$$b_T := \min_{\eta \in T_{\mathbf{u}^*}\mathcal{M} \setminus \{0\}} \frac{\langle B^*\eta, \eta \rangle_h}{\langle \eta, \eta \rangle_h} \geq \lambda_1(B^*), \quad \sigma_T := \min_{\eta \in T_{\mathbf{u}^*}\mathcal{M} \setminus \{0\}} \frac{\langle H^*\eta, \eta \rangle_h}{\langle \eta, \eta \rangle_h}, \quad (21)$$

for the smallest eigenvalues of the quadratic forms of B^* and of H^* on the tangent space.

LEMMA 3.2 (Coercivity of the constrained Hessian) Let $\Lambda := \max\{\lambda_2(A_\beta), \lambda_1(A_{3\beta})\}$ and $\sigma := \Lambda - \lambda_1(A_\beta)$. Then $\sigma_T \geq \sigma > 0$, and $b_T = \lambda_2(A_\beta)$ for the a_u metric.

Proof The tangent space is the $\langle \cdot, \cdot \rangle_h$ -orthogonal complement of $\mathbf{u}^*$, the eigenvector of the simple smallest eigenvalue of A_β , so the Courant–Fischer principle in $\langle \cdot, \cdot \rangle_h$ gives $\langle A_\beta \eta, \eta \rangle_h \geq \lambda_2(A_\beta) \langle \eta, \eta \rangle_h$ on $T_{\mathbf{u}^*}\mathcal{M}$, with equality for the second eigenvector, which is the claim on b_T . Writing $H^* = A_\beta - \lambda_1(A_\beta)\mathbb{I} + 2\beta \text{diag}((\mathbf{u}^*)^2)$ with $\text{diag}((\mathbf{u}^*)^2) \succeq 0$ gives $\sigma_T \geq \lambda_2(A_\beta) - \lambda_1(A_\beta)$, while $H^* \succeq \lambda_1(A_{3\beta}) - \lambda_1(A_\beta)$ on all of $\mathbb{R}^N$, so $\sigma_T \geq \sigma$. Finally $\lambda_1(A_{3\beta}) > \lambda_1(A_\beta)$. Let $\mathbf{v}$ be a normalized eigenvector of $A_{3\beta}$ for $\lambda_1(A_{3\beta})$. Since $A_{3\beta} - A_\beta = 2\beta \text{diag}((\mathbf{u}^*)^2) \succeq 0$, $\lambda_1(A_\beta) \leq \langle A_\beta \mathbf{v}, \mathbf{v} \rangle_h \leq \langle A_{3\beta} \mathbf{v}, \mathbf{v} \rangle_h = \lambda_1(A_{3\beta})$. If $\lambda_1(A_{3\beta}) = \lambda_1(A_\beta)$, then $\mathbf{v}$ attains the minimum of the Rayleigh quotient of A_β , so $\mathbf{v} = \pm \mathbf{u}^*$ by simplicity, and then $\lambda_1(A_{3\beta}) = \langle A_{3\beta} \mathbf{u}^*, \mathbf{u}^* \rangle_h = \lambda_1(A_\beta) + 2\beta \langle (\mathbf{u}^*)^2, (\mathbf{u}^*)^2 \rangle_h > \lambda_1(A_\beta)$, a contradiction. With $\lambda_2(A_\beta) > \lambda_1(A_\beta)$ this gives $\sigma > 0$. $\square$

LEMMA 3.3 (Discrete maximum principle) Under Assumption 2.1, $\beta \max_i (\mathbf{u}_i^*)^2 = \beta (\mathbf{u}_{i_0}^*)^2 \leq \lambda_1(A_\beta) - \mathbb{V}_{i_0} \leq \lambda_1(A_\beta)$, where i_0 is an index at which $\mathbf{u}_i^*$ is maximal.

Proof Since $\mathbf{u}^* > 0$ by Lemma 2.2, $\mathbb{S}_{i_0j} \leq 0$ and $\mathbf{u}_j^* \leq \mathbf{u}_{i_0}^*$ for $j \neq i_0$, and the row sums of $\mathbb{S}$ are nonnegative, $(\mathbb{S}\mathbf{u}^*)_{i_0} \geq (\sum_j \mathbb{S}_{i_0j})\mathbf{u}_{i_0}^* \geq 0$, hence $(-\Delta_h \mathbf{u}^*)_{i_0} \geq 0$, and the i_0 -th component of $A_\beta \mathbf{u}^* = \lambda_1(A_\beta) \mathbf{u}^*$ gives $\beta (\mathbf{u}_{i_0}^*)^2 \leq \lambda_1(A_\beta) - \mathbb{V}_{i_0}$, with $\mathbb{V} \geq 0$. $\square$

metric	$Z = H^* - B^*$	$\zeta_- \leq$	$b_T \geq$	$\zeta_+ \leq$ (Assumption 2.1)
modified H^1	$\mathbb{V} + 3\beta \text{diag}((\mathbf{u}^*)^2) - (\lambda_1 + \alpha)\mathbb{I}$	$\lambda_1 + \alpha$	$\lambda_1(-\Delta_h) + \alpha$	$(\max_i \mathbb{V}_i + 2\lambda_1 - \alpha)_+$
a_0	$3\beta \text{diag}((\mathbf{u}^*)^2) - \lambda_1 \mathbb{I}$	λ_1	$\lambda_1(-\Delta_h + \mathbb{V})$	$2\lambda_1$
a_u	$2\beta \text{diag}((\mathbf{u}^*)^2) - \lambda_1 \mathbb{I}$	λ_1	$\lambda_2(A_\beta)$	λ_1
modified a_0	$\mathbb{V}_2 + 3\beta \text{diag}((\mathbf{u}^*)^2) - \lambda_1 \mathbb{I}$	$\lambda_1 + \mathcal{V}_-$	$\lambda_1(A_1)$	$(\mathcal{V}_+ + 2\lambda_1)_+$
modified a_u	$-\lambda_1 \mathbb{I}$	λ_1	$\lambda_1 + \sigma_T$	0
Thomas–Fermi	Z_{TF} of (26)	$e_{\text{TF}} + \ell \lambda_1^{1/3}$	$\ell \lambda_1^{1/3}$	$(e_{\text{TF}} - \ell \lambda_1^{1/3})_+$

TABLE 2 *The difference $Z = H^* - B^*$ at the ground state and bounds on $\zeta_\pm$ of (20) and b_T of (21). Here $\lambda_1 = \lambda_1(A_\beta)$, $\mathcal{V}_\pm := \max_i (\pm(\mathbb{V}_2)_{ii})_+$ and e_{TF} is the Thomas–Fermi error (27). The last column uses Lemma 3.3, and the Thomas–Fermi entries hold without Assumption 2.1. The entries b_T of the a_u and modified a_u metrics and $\zeta_\pm$ of the modified a_u metric are exact.*

THEOREM 3.4 (Two-sided bounds of the Rayleigh quotient) Let $\mathbf{u}^*$ be the ground state. For every metric and every $\eta \in T_{\mathbf{u}^*}\mathcal{M} \setminus \{0\}$,

$$\frac{\sigma_T}{\sigma_T + \zeta_-} \leq \rho_{\mathbf{u}^*}(\eta) \leq 1 + \frac{\zeta_+}{b_T}, \quad \text{thus} \quad \kappa(\mathbf{u}^*) \leq \left(1 + \frac{\zeta_+}{b_T}\right) \left(1 + \frac{\zeta_-}{\sigma}\right) \quad (22)$$

for the condition number $\kappa(\mathbf{u}^*)$ of $\text{Hess} E(\mathbf{u}^*)$, the ratio of its largest to its smallest eigenvalue.

Proof Since Z and $\mathbb{M}$ are diagonal and $\mathbb{M}$ is positive, $-\zeta_- \langle \eta, \eta \rangle_h \leq \langle Z\eta, \eta \rangle_h \leq \zeta_+ \langle \eta, \eta \rangle_h$. On $T_{\mathbf{u}^*}\mathcal{M}$ therefore $H^* = B^* + Z \preceq B^* + \zeta_+ \mathbb{I} \preceq (1 + \zeta_+/b_T)B^*$ and $B^* = H^* - Z \preceq H^* + \zeta_- \mathbb{I} \preceq (1 + \zeta_-/\sigma_T)H^*$ as quadratic forms, which by (19) are the two bounds in (22). The bound on $\kappa(\mathbf{u}^*)$ follows from $\sigma_T \geq \sigma$. $\square$

COROLLARY 3.5 (Condition numbers of the classical metrics) Let $\mathbf{u}^*$ be the ground state, let Λ be as in Lemma 3.2, and let Assumption 2.1 hold. Then

$$\begin{aligned}\kappa_{a_u}(\mathbf{u}^*) &\leq \left(1 + \frac{\zeta_+}{\lambda_2(A_\beta)}\right) \frac{\Lambda}{\Lambda - \lambda_1(A_\beta)} \leq \left(1 + \frac{\lambda_1(A_\beta)}{\lambda_2(A_\beta)}\right) \frac{\Lambda}{\Lambda - \lambda_1(A_\beta)} \leq \frac{2\lambda_1(A_{3\beta})}{\lambda_1(A_{3\beta}) - \lambda_1(A_\beta)}, \\ \kappa_{a_0}(\mathbf{u}^*) &\leq \left(1 + \frac{2\lambda_1(A_\beta)}{\lambda_1(-\Delta_h + \mathbb{V})}\right) \frac{\Lambda}{\Lambda - \lambda_1(A_\beta)}, \\ \kappa_{H^1}(\mathbf{u}^*) &\leq \left(1 + \frac{(\max_i \mathbb{V}_i + 2\lambda_1(A_\beta) - \alpha)_+}{\lambda_1(-\Delta_h) + \alpha}\right) \frac{\Lambda + \alpha}{\Lambda - \lambda_1(A_\beta)},\end{aligned}$$

and for the a_u metric $1 - \lambda_1(A_\beta)/\Lambda \leq \rho_{\mathbf{u}^*}(\eta) \leq 1 + \lambda_1(A_\beta)/\lambda_2(A_\beta) < 2$ on $T_{\mathbf{u}^*}\mathcal{M} \setminus \{0\}$, with $\zeta_+ = 0$ and the spectrum in $[1 - \lambda_1(A_\beta)/\Lambda, 1]$ as soon as $2\beta \max_i (\mathbf{u}_i^*)^2 \leq \lambda_1(A_\beta)$. Without Assumption 2.1, the same bounds hold with $\zeta_+ = \max_i (Z_{ii})_+$ of Table 2 in place of the numerators $\lambda_1(A_\beta)$, $2\lambda_1(A_\beta)$ and $(\max_i \mathbb{V}_i + 2\lambda_1(A_\beta) - \alpha)_+$, and the lower bound $1 - \lambda_1(A_\beta)/\Lambda$ is unchanged. In particular, the bound $\rho_{\mathbf{u}^*}(\eta) < 2$ for the a_u metric becomes $\rho_{\mathbf{u}^*}(\eta) \leq 1 + \zeta_+/\lambda_2(A_\beta)$, and the factor 2 in the last bound on $\kappa_{a_u}(\mathbf{u}^*)$ becomes $1 + \zeta_+/\lambda_2(A_\beta)$.

Proof The matrices Z in Table 2 are read off (10). The bounds on ζ_- follow from $(\mathbf{u}_i^*)^2 \geq 0$ and $\mathbb{V} \geq 0$, the bounds on b_T follow from (21) and Lemma 3.2, and the bounds on ζ_+ follow from Lemma 3.3, for instance $2\beta(\mathbf{u}_i^*)^2 - \lambda_1(A_\beta) \leq \lambda_1(A_\beta) - 2\mathbb{V}_{i_0} \leq \lambda_1(A_\beta)$ and $3\beta(\mathbf{u}_i^*)^2 - \lambda_1(A_\beta) \leq 2\lambda_1(A_\beta)$. Theorem 3.4 gives the bounds, with $\sigma/(\sigma + \lambda_1(A_\beta)) = 1 - \lambda_1(A_\beta)/\Lambda$ and $1 + (\lambda_1(A_\beta) + \alpha)/\sigma = (\Lambda + \alpha)/(\Lambda - \lambda_1(A_\beta))$ by $\sigma = \Lambda - \lambda_1(A_\beta)$. The last inequality for $\kappa_{a_u}(\mathbf{u}^*)$ follows from $1 + \lambda_1(A_\beta)/\lambda_2(A_\beta) < 2$ and from the fact that $\Lambda/(\Lambda - \lambda_1(A_\beta))$ decreases in $\Lambda \geq \lambda_1(A_{3\beta})$. $\square$

For the a_u metric in the harmonic trap $V = |\mathbf{x}|^2$, Appendix A compares these bounds with the exact extreme eigenvalues of $\text{Hess}E(\mathbf{u}^*)$ in one, two and three dimensions. For $\beta \geq 10^2$, the lower bound $1 - \lambda_1(A_\beta)/\Lambda$ is within 9% of the exact minimum in one dimension and within 21% in two and three dimensions. The bound $1 - \lambda_1(A_\beta)/\lambda_2(A_\beta)$ based on the spectral gap alone is smaller than the exact minimum by a factor up to 3×10^4 .

3.3. The modified a_0 metric

Consider a splitting $V = V_1 + V_2$ as in Section 2.2, where $V_1 \geq 0$ is separable, $V_1(\mathbf{x}) = \sum_{j=1}^d v_j(x_j)$, and V_2 is bounded. Then the operator $A_1 = -\Delta_h + \mathbb{V}_1$ can be inverted exactly and efficiently on GPU by tensor-product direct solvers (Liu and Zhang, 2026). When V_1 is confining, that is, $V_1(\mathbf{x}) \rightarrow \infty$ as $|\mathbf{x}| \rightarrow \infty$, the preconditioned operator $A_1^{-1}(-\Delta_h + \mathbb{V}) = \mathbb{I} + A_1^{-1}\mathbb{V}_2$ has eigenvalues clustered at one and a condition number bounded independently of the mesh size and the domain size (Liu and Zhang, 2026, Theorems 3.4 and 3.5). This motivates taking A_1 as the metric operator itself, which defines the modified a_0 metric (10). By (19) the Rayleigh quotient of the Riemannian Hessian for this metric at the ground state is $1 + \langle Z\eta, \eta \rangle_h / \langle A_1\eta, \eta \rangle_h$ with $Z = \mathbb{V}_2 + 3\beta \text{diag}((\mathbf{u}^*)^2) - \lambda_1(A_\beta)\mathbb{I}$, which is the matrix Z of the a_0 metric plus the nonseparable part of the potential.

PROPOSITION 3.6 (Condition number of the modified a_0 metric) Let $\mathbf{u}^*$ be the ground state, Λ as in Lemma 3.2 and $\mathcal{V}_\pm = \max_i (\pm(\mathbb{V}_2)_{ii})_+$. Under the modified a_0 metric, denoted by $\tilde{a}_0$ in subscripts, for

all $\eta \in T_{\mathbf{u}^*} \mathcal{M} \setminus \{0\}$,

$$\frac{\Lambda - \lambda_1(A_\beta)}{\Lambda + \mathcal{V}_-} \leq \rho_{\mathbf{u}^*}(\eta) \leq 1 + \frac{\max_i ((\mathbb{V}_2)_{ii} + 3\beta(\mathbf{u}_i^*)^2 - \lambda_1(A_\beta))_+}{\lambda_1(A_1)},$$

and under Assumption 2.1

$$\kappa_{\tilde{a}_0}(\mathbf{u}^*) \leq \left(1 + \frac{(\mathcal{V}_+ + 2\lambda_1(A_\beta))_+}{\lambda_1(A_1)}\right) \frac{\Lambda + \mathcal{V}_-}{\Lambda - \lambda_1(A_\beta)}.$$

If $V_2 \geq 0$, then $\kappa_{\tilde{a}_0}(\mathbf{u}^*) \leq (1 + \mathcal{V}_+/\lambda_1(A_1)) \kappa_{a_0}(\mathbf{u}^*)$.

Proof The entries of $Z = \mathbb{V}_2 + 3\beta \text{diag}((\mathbf{u}^*)^2) - \lambda_1(A_\beta)\mathbb{I}$ are at least $-\mathcal{V}_- - \lambda_1(A_\beta)$ and, for a monotone scheme, at most $\mathcal{V}_+ + 2\lambda_1(A_\beta)$ by Lemma 3.3, and $b_T \geq \lambda_1(A_1)$ by (21), so Theorem 3.4 with $\sigma = \Lambda - \lambda_1(A_\beta)$ gives the first two claims. For the last claim, $-\Delta_h + \mathbb{V} = A_1 + \mathbb{V}_2$ and $0 \leq \mathbb{V}_2 \leq \mathcal{V}_+\mathbb{I}$ give $\langle A_1 \mathbf{x}, \mathbf{x} \rangle_h \leq \langle (A_1 + \mathbb{V}_2) \mathbf{x}, \mathbf{x} \rangle_h \leq (1 + \mathcal{V}_+/\lambda_1(A_1)) \langle A_1 \mathbf{x}, \mathbf{x} \rangle_h$, so by (19) the Rayleigh quotients of the two metrics differ pointwise by a factor between 1 and $1 + \mathcal{V}_+/\lambda_1(A_1)$, and so do their extremes. $\square$

Compared with the bound on $\kappa_{a_0}(\mathbf{u}^*)$ in Corollary 3.5, the bound in Proposition 3.6 replaces $\lambda_1(-\Delta_h + \mathbb{V})$ by $\lambda_1(A_1)$ and adds the bounded quantities $\mathcal{V}_\pm$ to the numerators. Thus the modified a_0 metric keeps the conditioning of the a_0 metric up to a bounded factor. Unlike the modified H^1 metric, whose bound grows with the domain size, it is robust with respect to the domain size.

REMARK 3.7 (Domain robustness of the modified a_0 metric) Let Assumption 2.1 hold, let V_1 be confining and V_2 be bounded on $\mathbb{R}^d$, and let the problem be posed on the boxes $\Omega_L = (-L, L)^d$ with a fixed mesh size h and $L \in h\mathbb{N}$, so that the grids are nested. Then $\lambda_1(A_1)$ is bounded below uniformly in L , because the first eigenvalue of $-\Delta_h + \mathbb{V}_1$ on Ω_L is at least that on the infinite grid, by the min-max argument of Liu and Zhang (2026, proof of Theorem 3.5). By (12), $\lambda_1(A_\beta) \leq 4E(\mathbf{u}^*)$, which is bounded above uniformly in L by comparison with a fixed grid function supported in the smallest box. Thus, if the coercivity constant $\sigma = \Lambda - \lambda_1(A_\beta)$ is bounded below uniformly in L , then $\kappa_{\tilde{a}_0}(\mathbf{u}^*)$ is bounded uniformly in L by Proposition 3.6, since $(\Lambda + \mathcal{V}_-)/(\Lambda - \lambda_1(A_\beta)) = 1 + (\lambda_1(A_\beta) + \mathcal{V}_-)/\sigma$. For the modified H^1 metric, the bound in Corollary 3.5 grows with L , because $\lambda_1(-\Delta_h) \rightarrow 0$ and $\max_i \mathbb{V}_i \rightarrow \infty$ as $L \rightarrow \infty$. In the tests of Section 5.4, the iteration count of the modified H^1 flow grows with L while that of the modified a_0 flow does not.

3.4. The modified a_u metric

On the tangent space, $g_{\mathbf{u}^*}(\text{Hess} E(\mathbf{u}^*)[\eta], \eta) = g_{\mathbf{u}^*}(\eta, \eta) + \langle Z\eta, \eta \rangle_h$ by (19), so the closer the metric operator is to the constrained Hessian $H^* = A_{3\beta} - \lambda_1(A_\beta)\mathbb{I}$, the closer the Riemannian Hessian is to the identity. A metric operator must be positive definite at every $\mathbf{u} \in \mathbb{R}^N$. A natural choice is $A_{3\beta}(\mathbf{u})$ itself, which is positive definite for every $\mathbf{u}$ and at $\mathbf{u}^*$ differs from H^* only by the scalar shift $\lambda_1(A_\beta)\mathbb{I}$. Including the shift, which amounts to Newton's method, destroys positive definiteness away from the ground state. The operator $A_{3\beta}(\mathbf{u})$ thus defines the modified a_u metric (10). At the ground state, $Z = -\lambda_1(A_\beta)\mathbb{I}$ is a scalar matrix, so the Riemannian Hessian is explicit.

PROPOSITION 3.8 (Condition number of the modified a_u metric) Let $\mathbf{u}^*$ be the ground state and let $s_1 \leq \dots \leq s_{N-1}$ be the eigenvalues of the quadratic form $\langle H^* \eta, \eta \rangle_h$ on $T_{\mathbf{u}^*} \mathcal{M}$, that is, of the $\langle \cdot, \cdot \rangle_h$ -orthogonal compression of H^* to the tangent space, so that $s_1 = \sigma_T$. Under the modified a_u metric, denoted by $\tilde{a}_u$ in subscripts, the eigenvalues of $\text{Hess} E(\mathbf{u}^*)$ are

$$\rho_k = \frac{s_k}{s_k + \lambda_1(A_\beta)}, \quad k = 1, \dots, N-1, \quad (23)$$

and thus, without any assumption on the scheme, $\sigma_T / (\lambda_1(A_\beta) + \sigma_T) \leq \rho_{\mathbf{u}^*}(\eta) < 1$ on $T_{\mathbf{u}^*} \mathcal{M} \setminus \{0\}$ and

$$\kappa_{\tilde{a}_u}(\mathbf{u}^*) = \frac{s_{N-1}}{s_{N-1} + \lambda_1(A_\beta)} \left(1 + \frac{\lambda_1(A_\beta)}{\sigma_T} \right) \leq 1 + \frac{\lambda_1(A_\beta)}{\sigma_T}.$$

Proof Since $\text{Hess} E(\mathbf{u}^*)$ is self-adjoint with respect to $g_{\mathbf{u}^*}$, polarization of (18) (Absil et al., 2008, (5.36)) gives $g_{\mathbf{u}^*}(\text{Hess} E(\mathbf{u}^*)[\eta], \xi) = \langle H^* \eta, \xi \rangle_h$ for all $\eta, \xi \in T_{\mathbf{u}^*} \mathcal{M}$. Hence $\text{Hess} E(\mathbf{u}^*)[\eta] = \rho \eta$ if and only if $\langle H^* \eta, \xi \rangle_h = \rho \langle B^* \eta, \xi \rangle_h$ for all $\xi \in T_{\mathbf{u}^*} \mathcal{M}$. With $B^* = A_{3\beta} = H^* + \lambda_1(A_\beta)\mathbb{I}$ this reads $(1 - \rho) \langle H^* \eta, \xi \rangle_h = \rho \lambda_1(A_\beta) \langle \eta, \xi \rangle_h$. Here $\rho \neq 1$, since $\rho = 1$ forces $\eta = 0$. Thus η is an eigenvector of $\text{Hess} E(\mathbf{u}^*)$ with eigenvalue ρ if and only if it is an eigenvector of the compression of H^* to $T_{\mathbf{u}^*} \mathcal{M}$ with eigenvalue $s = \rho \lambda_1(A_\beta) / (1 - \rho)$, that is, $\rho = s / (s + \lambda_1(A_\beta))$. $\square$

Unlike for the a_u metric, $\nabla_X E(\mathbf{u}) = \mathbf{u} - 2\beta A_{3\beta}(\mathbf{u})^{-1} \mathbf{u}^3$ is not $\mathbf{u}$, so one iteration requires two solves instead of one. Remark 3.12 compares this cost with the gain in conditioning in the Thomas–Fermi regime.

REMARK 3.9 (Family of metrics between a_u and modified a_u) The a_u and modified a_u metrics are the members $c = 1$ and $c = 3$ of the family of $\mathbf{u}$ -dependent metrics induced by

$$B_c(\mathbf{u}) := -\Delta_h + \mathbb{V} + c\beta \text{diag}(\mathbf{u}^2), \quad c \geq 1, \quad (24)$$

for which $Z_c = (3 - c)\beta \text{diag}((\mathbf{u}^*)^2) - \lambda_1(A_\beta)\mathbb{I}$ at the ground state. For $c \leq 3$ the entries of Z_c are at least $-\lambda_1(A_\beta)$, so $\zeta_- \leq \lambda_1(A_\beta)$, and under Assumption 2.1 they are at most $(2 - c)\lambda_1(A_\beta)$ by Lemma 3.3, so $\zeta_+ = 0$ for every $c \in [2, 3]$. Thus, under Assumption 2.1, all members with $c \in [2, 3]$ share the bound $\kappa_c(\mathbf{u}^*) \leq 1 + \lambda_1(A_\beta) / \sigma_T$ of Proposition 3.8. Only $c = 3$ has a scalar Z_c and thus the explicit spectrum (23), and for this member the bound holds without any assumption on the scheme (Proposition 3.8).

3.5. The Thomas–Fermi metric

We now explain how the new Thomas–Fermi metric is designed. By the Rayleigh quotient (19), the metric operator should approximate the constrained Hessian $H^* = A_{3\beta} - \lambda_1\mathbb{I}$ on the tangent space, where $\lambda_1 := \lambda_1(A_\beta)$. The operator $A_{3\beta}(\mathbf{u}) - \lambda(\mathbf{u})\mathbb{I}$ at a general iterate $\mathbf{u}$ is not positive definite (Section 3.4), so H^* cannot be used directly. Inspired by the Thomas–Fermi approximation $\beta(\mathbf{u}^*)^2 \approx (\lambda_1 - \mathbb{V})_+$, if we write $\mathbb{V} - \lambda_1\mathbb{I} = (\mathbb{V} - \lambda_1)_+ - (\lambda_1 - \mathbb{V})_+$, then

$$H^* = \underbrace{-\Delta_h + 2\beta \text{diag}((\mathbf{u}^*)^2) + \text{diag}((\mathbb{V} - \lambda_1)_+)}_{\succ 0} + \text{diag}(\beta(\mathbf{u}^*)^2 - (\lambda_1 - \mathbb{V})_+), \quad (25)$$

where the last term is the error of the Thomas–Fermi approximation. The first term is positive definite, and it remains so at a general iterate $\mathbf{u}$ with $\lambda(\mathbf{u})$ of (12) in place of λ_1 , because each of its three

terms is positive semidefinite. Replacing the Thomas–Fermi approximation error by $\ell\lambda(\mathbf{u})^{1/3}\mathbb{I}$ gives the Thomas–Fermi metric operator $B_{\text{TF}}(\mathbf{u})$ of (11), where the power $1/3$ is taken according to the analysis in Gallo and Pelinovsky (2011) and Corollary 3.11 below. To our knowledge, the metric (11) has not previously been proposed. The Thomas–Fermi preconditioner of Antoine and Duboscq (2014), used alone or combined with a Laplace preconditioner in Antoine et al. (2017), is the diagonal operator $\alpha + \mathbb{V} + \beta\mathbf{u}^2$ obtained by dropping the kinetic term as in the Thomas–Fermi approximation, whereas the metric here keeps the Laplacian.

At the ground state, $\lambda(\mathbf{u}^*) = \lambda_1$ and the matrix Z of the metric is the Thomas–Fermi error shifted down by $\ell\lambda_1^{1/3}$,

$$Z_{\text{TF}} = \text{diag}(\beta(\mathbf{u}^*)^2 - (\lambda_1 - \mathbb{V})_+) - \ell\lambda_1^{1/3}\mathbb{I}. \quad (26)$$

PROPOSITION 3.10 (Condition number of the Thomas–Fermi metric) Let $\mathbf{u}^*$ be the ground state, $\lambda_1 = \lambda_1(A_\beta)$ and

$$e_{\text{TF}} := \max_i |\beta(\mathbf{u}_i^*)^2 - (\lambda_1 - \mathbb{V}_i)_+|. \quad (27)$$

Under the Thomas–Fermi metric with any $\ell > 0$, denoted by TF in subscripts, the spectrum of $\text{Hess}E(\mathbf{u}^*)$ lies in $[\sigma_T/(\sigma_T + e_{\text{TF}} + \ell\lambda_1^{1/3}), \max\{1, e_{\text{TF}}/(\ell\lambda_1^{1/3})\}]$, and thus

$$\kappa_{\text{TF}}(\mathbf{u}^*) \leq \max\left\{1, \frac{e_{\text{TF}}}{\ell\lambda_1^{1/3}}\right\} \left(1 + \frac{e_{\text{TF}} + \ell\lambda_1^{1/3}}{\sigma_T}\right). \quad (28)$$

In particular, if $e_{\text{TF}} \leq C_{\text{TF}}\lambda_1^{1/3}$ and $\sigma_T \geq c_\sigma\lambda_1^{1/3}$ with constants independent of β , then

$$\kappa_{\text{TF}}(\mathbf{u}^*) \leq \max\left\{1, \frac{C_{\text{TF}}}{\ell}\right\} \left(1 + \frac{C_{\text{TF}} + \ell}{c_\sigma}\right) \quad (29)$$

uniformly in β , for every $\ell > 0$.

Proof Every diagonal entry of (26) lies in $[-e_{\text{TF}} - \ell\lambda_1^{1/3}, e_{\text{TF}} - \ell\lambda_1^{1/3}]$, so $\zeta_- \leq e_{\text{TF}} + \ell\lambda_1^{1/3}$ and $\zeta_+ \leq (e_{\text{TF}} - \ell\lambda_1^{1/3})_+$, and $b_T \geq \lambda_1(B^*) \geq \ell\lambda_1^{1/3}$ by (21), because $B_{\text{TF}}(\mathbf{u}^*) - \ell\lambda_1^{1/3}\mathbb{I}$ is positive semidefinite. Theorem 3.4 gives the spectrum, since $1 + \zeta_+/b_T \leq 1 + (e_{\text{TF}}/(\ell\lambda_1^{1/3}) - 1)_+ = \max\{1, e_{\text{TF}}/(\ell\lambda_1^{1/3})\}$, and the last claim follows by inserting the stated hypotheses. $\square$

For strong interaction, the following result, which is a simple corollary of the results in Gallo and Pelinovsky (2011), gives the sharp deterioration of the a_u metric for the one-dimensional harmonic trap.

COROLLARY 3.11 (Thomas–Fermi asymptotics of the coercivity constant) For the continuous problem on $\Omega = \mathbb{R}$ with $V(x) = x^2$, let u be the positive minimizer of $\mathcal{E}$ on $\{u \in H^1(\mathbb{R}) : \int_{\mathbb{R}} x^2 u^2 dx < \infty, \|u\|_{L^2(\mathbb{R})} = 1\}$, let $A_{c\beta} := -\partial_x^2 + x^2 + c\beta u^2$ on $L^2(\mathbb{R})$, $T_u := \{\eta \in H^1(\mathbb{R}) : x\eta \in L^2(\mathbb{R}), (\eta, u) = 0\}$, and let $\lambda_1 = \lambda_1(A_\beta)$ be the eigenvalue of u , which tends to infinity as $\beta \rightarrow \infty$. Then, as $\beta \rightarrow \infty$,

$$\lambda_1(A_{3\beta}) - \lambda_1(A_\beta) = \mu_1 \lambda_1^{1/3} (1 + o(1)), \quad \mu_1 = 2.4105\dots, \quad (30)$$

where μ_1 is the smallest eigenvalue of the Schrödinger operator $M_0 = -4\partial_y^2 + 3v_0(y)^2 - y$ on $L^2(\mathbb{R})$ and v_0 is the Hastings–McLeod solution of the Painlevé-II equation $4v'' + yv - v^3 = 0$, characterized

by $v_0(y) \sim \sqrt{y}$ as $y \rightarrow +\infty$ and $v_0(y) \rightarrow 0$ as $y \rightarrow -\infty$. Thus, for the a_u metric,

$$\min_{T_u} \rho_u = \mu_1 \lambda_1^{-2/3} (1 + o(1)), \quad \max_{T_u} \rho_u \rightarrow 2, \quad \kappa_{a_u}(u) = \frac{2}{\mu_1} \lambda_1^{2/3} (1 + o(1)), \quad (31)$$

and the lower bound $1 - \lambda_1(A_\beta)/\Lambda$ of Corollary 3.5 is asymptotically exact.

Proof Equation (30) is Gallo and Pelinovsky (2011, Theorem 2) in our variables. Set $x = \sqrt{\lambda_1}y$ and $u(x) = \sqrt{\lambda_1/\beta} U(y)$. The ground state equation becomes $\varepsilon^2 U'' + (1 - y^2)U - U^3 = 0$ with $\varepsilon = \lambda_1^{-1}$, and the continuous constrained Hessian $H := A_\beta - \lambda_1$ satisfies $\lambda_1^{-1}H = -\varepsilon^2 \partial_y^2 + 3U^2 - 1 + y^2 = L_+^\varepsilon$ in the notation of Gallo and Pelinovsky (2011). By Gallo and Pelinovsky (2011, Theorem 2), the first two eigenvalues $\delta_1 \leq \delta_2$ of H are

$$\delta_j = \mu_1 \lambda_1^{1/3} (1 + o(1)), \quad j = 1, 2.$$

These two eigenvalues correspond to the two boundary layers, and $\mu_1 > 0$ by Gallo and Pelinovsky (2011, Lemma 2.2). Since $\delta_1 = \lambda_1(A_{3\beta}) - \lambda_1$, this proves (30).

As quadratic forms, $\lambda_1 \mathbb{I} \preceq A_\beta \preceq H + \lambda_1 \mathbb{I}$, so $\langle H\eta, \eta \rangle / \langle A_\beta \eta, \eta \rangle \geq \langle H\eta, \eta \rangle / (\langle H\eta, \eta \rangle + \lambda_1 \|\eta\|^2) \geq \delta_1 / (\delta_1 + \lambda_1)$ on T_u , because $t \mapsto t/(t + \lambda_1)$ is increasing and $\langle H\eta, \eta \rangle \geq \delta_1 \|\eta\|^2$. The second eigenfunction v_2 of the even Sturm–Liouville operator H is odd, hence orthogonal to u , and it gives the upper bound in

$$\frac{\delta_1}{\delta_1 + \lambda_1} \leq \min_{T_u} \frac{\langle H\eta, \eta \rangle}{\langle A_\beta \eta, \eta \rangle} \leq \frac{\langle Hv_2, v_2 \rangle}{\lambda_1 \|v_2\|^2} = \frac{\delta_2}{\lambda_1}.$$

Both endpoints are $\mu_1 \lambda_1^{-2/3} (1 + o(1))$. The bound $\sigma/(\sigma + \lambda_1) = 1 - \lambda_1/\Lambda$ lies between the left endpoint and the minimum, since $\sigma \geq \delta_1$, so it too is asymptotically exact. Finally, $\rho_u(\eta) = 1 + \langle (2\beta u^2 - \lambda_1)\eta, \eta \rangle / \langle A_\beta \eta, \eta \rangle \leq 2$, because $\beta u^2 \leq \lambda_1$ by the maximum principle at the maximum of u and $A_\beta \succeq \lambda_1 \mathbb{I}$. A fixed odd bump φ supported near the origin gives $\rho_u(\varphi) \rightarrow 2$, because $\beta u^2 = \lambda_1 - x^2 + o(\lambda_1)$ uniformly there by Gallo and Pelinovsky (2011, Theorem 1). This proves (31). $\square$

REMARK 3.12 (Condition numbers in the Thomas–Fermi regime) For the continuous one-dimensional harmonic problem of Corollary 3.11, the condition numbers behave as follows as $\beta \rightarrow \infty$, with $\lambda_1 = \lambda_1(A_\beta) \rightarrow \infty$.

- *The a_u metric.* By (31), $\kappa_{a_u}(u) = (2/\mu_1) \lambda_1^{2/3} (1 + o(1))$ grows like $\lambda_1^{2/3}$.
- *The family (24).* Since $A_\beta \preceq B_c \preceq cA_\beta$ for $c \geq 1$, (19) gives $\kappa_c(u) \geq \kappa_{a_u}(u)/c$, so every member grows at least like $\lambda_1^{2/3}$. Its member $c = 3$, the modified a_u metric, has $\kappa_{\tilde{a}_u}(u) = 1 + \lambda_1/\sigma_T \sim \mu_1^{-1} \lambda_1^{2/3}$ by Proposition 3.8 and (30), since the Rayleigh quotient $s/(s + \lambda_1)$ in (23) has supremum 1 over the unbounded spectrum of H on T_u , approached by high modes and not attained. This is half of $\kappa_{a_u}(u)$, but each iteration needs two solves instead of one (Section 3.4).
- *The a_0 metric.* Corollary 3.5 with $\lambda_1(-\partial_x^2 + x^2) = 1$ and $\sigma \geq \lambda_1(A_{3\beta}) - \lambda_1$ gives $\kappa_{a_0}(u) \leq (1 + 2\lambda_1)(1 + \lambda_1/(\lambda_1(A_{3\beta}) - \lambda_1)) = (2/\mu_1) \lambda_1^{5/3} (1 + o(1))$, so the bound grows like $\lambda_1^{5/3}$.
- *The Thomas–Fermi metric.* Its condition number is bounded uniformly in β for every $\ell > 0$, because both hypotheses of (29) hold for large β . The two eigenvalues in the proof of Corollary 3.11 enclose

σ_T , so $\sigma_T = \mu_1 \lambda_1^{1/3} (1 + o(1))$ and any $c_\sigma < \mu_1$ works. The uniform approximation of βu^2 by $\lambda_1^{1/3} v_0(y)^2$ in the layer variable $y = (\lambda_1 - x^2)/\lambda_1^{1/3}$ (Gallo and Pelinovsky, 2011, Theorem 1 and (3.17)) gives $e_{\text{TF}} = (C_0 + o(1)) \lambda_1^{1/3}$ with $C_0 = \sup_y |v_0(y)^2 - y_+|$, which is finite by Gallo and Pelinovsky (2011, Proposition 2.1), so any $C_{\text{TF}} > C_0$ works. Numerically $C_0 = v_0(0)^2 \approx 0.43$, so C_{TF} can be taken in $(C_0, 1/2)$, and for $\ell = 1/2$ the first factor in (29) equals one. Unlike the a_u metric and its family, the Thomas–Fermi metric is not uniformly equivalent to A_β , since on the modes that attain σ_T its quadratic form is of order $\lambda_1^{1/3}$ while that of A_β is of order λ_1 .

Appendix A tests these orders, in particular the uniform bound for the Thomas–Fermi metric, against exact spectra of the Riemannian Hessian in one and three dimensions.

4. Convergence

The results of this section are standard in form. A local linear rate determined by the spectrum of the Riemannian Hessian and global convergence to a critical point are classical for gradient descent on manifolds (Absil et al., 2008; Boumal, 2023; Schneider and Uschmajew, 2015). The point of this section is a uniform treatment of all the Sobolev gradient flows of Table 1 under the same hypotheses. For instance, it is not clear whether convergence results for one specific flow, such as those for the discrete modified H^1 flow in Chen et al. (2025), extend to the Thomas–Fermi metric, whose operator is continuous but not smooth in $\mathbf{u}$. Throughout this section we fix one of the six metrics, write $\|\eta\|_{\mathbf{u}} := g_{\mathbf{u}}(\eta, \eta)^{1/2}$ for the metric norm, and use the following uniform bounds.

LEMMA 4.1 (Uniform bounds on the metric and the energy) For a fixed discretization and fixed parameters, there are constants $0 < c_1 \leq c_2 < \infty$ such that $c_1 \|\mathbf{x}\|_h^2 \leq \|\mathbf{x}\|_{\mathbf{u}}^2 \leq c_2 \|\mathbf{x}\|_h^2$ for all $\mathbf{u} \in \mathcal{M}$ and $\mathbf{x} \in \mathbb{R}^N$. Thus $K := \{(\mathbf{u}, \eta) : \mathbf{u} \in \mathcal{M}, \eta \in T_{\mathbf{u}}\mathcal{M}, \|\eta\|_{\mathbf{u}} = 1\}$ is compact, and $C_g := \sup_{\mathbf{u} \in \mathcal{M}} \|\text{grad} E(\mathbf{u})\|_{\mathbf{u}}$ and $C_E := \max\{1, \sup\{|\frac{d^2}{dt^2} E(R_{\mathbf{u}}(t\eta))| : (\mathbf{u}, \eta) \in K, |t| \leq C_g\}\}$ are finite.

Proof The function $(\mathbf{u}, \mathbf{x}) \mapsto g_{\mathbf{u}}(\mathbf{x}, \mathbf{x})$ is positive and continuous on the compact set $\mathcal{M} \times \{\mathbf{x} : \|\mathbf{x}\|_h = 1\}$. For the Thomas–Fermi metric, the continuity holds because $\lambda(\mathbf{u})$ is continuous. We take c_1 and c_2 to be the minimum and the maximum of this function. The set K is closed. Since $\|\eta\|_h \leq c_1^{-1/2}$ for $(\mathbf{u}, \eta) \in K$, the set K is also bounded and therefore compact. The inverse $B(\mathbf{u})^{-1}$ is positive definite and depends continuously on $\mathbf{u}$. By (15), $\text{grad} E$ is therefore continuous on $\mathcal{M}$. Since $\mathcal{M}$ is compact and the metric is continuous, C_g is finite.

For $(\mathbf{u}, \eta) \in K$, we have $R_{\mathbf{u}}(t\eta) = (\mathbf{u} + t\eta)/(1 + t^2 \|\eta\|_h^2)^{1/2}$ by (16), and the right-hand side is a smooth function of $(\mathbf{u}, \eta, t) \in \mathbb{R}^N \times \mathbb{R}^N \times \mathbb{R}$. Since E is a polynomial, the second derivative of $E(R_{\mathbf{u}}(t\eta))$ in t is therefore a continuous function of $(\mathbf{u}, \eta, t)$. Hence C_E is finite, because $K \times [-C_g, C_g]$ is compact. Similar bounds for smooth metrics are given in Boumal (2023, Lemma 10.57). $\square$

4.1. Local convergence via the Riemannian Hessian

Let $\mathbf{u}^*$ be the ground state. As in Section 3.2, we assume that $\mathbf{u}^*$ is an eigenvector of the simple smallest eigenvalue $\lambda_1(A_\beta)$ of A_β , which holds under Assumption 2.1. For the Thomas–Fermi metric, Theorem 4.2 also assumes that no node satisfies $\nabla_i = \lambda_1(A_\beta)$, so that the metric is smooth near $\mathbf{u}^*$ (Appendix B). Proposition 4.3 does not need this assumption. By (19) and Lemma 3.2, $\text{Hess} E(\mathbf{u}^*)$ is positive definite.

We denote the eigenvalues of $\text{Hess}E(\mathbf{u}^*)$ by $0 < \rho_1 \leq \dots \leq \rho_{N-1}$. Section 3 bounds the condition number $\kappa(\mathbf{u}^*) = \rho_{N-1}/\rho_1$ for each metric. The local rate of the iteration (16) with a fixed step is controlled by this spectrum. Boumal (2023, Theorem 4.20) proves local linear convergence for every fixed step $\tau < 2/\rho_{N-1}$. Theorem 4.2 gives the rate in terms of the spectrum and identifies the optimal step.

THEOREM 4.2 (Local rate of the fixed-step iteration) For $\tau > 0$ let $F_\tau(\mathbf{u}) := R_{\mathbf{u}}(-\tau \text{grad}E(\mathbf{u}))$, so that (16) with the constant step τ reads $\mathbf{u}_{k+1} = F_\tau(\mathbf{u}_k)$. Then F_τ is smooth near $\mathbf{u}^*$, $F_\tau(\mathbf{u}^*) = \mathbf{u}^*$, and

$$DF_\tau(\mathbf{u}^*)[\xi] = \xi - \tau \text{Hess}E(\mathbf{u}^*)[\xi] \quad \text{for all } \xi \in T_{\mathbf{u}^*}\mathcal{M}. \quad (32)$$

Thus, if $0 < \tau < 2/\rho_{N-1}$, then for every $\varepsilon > 0$ there are a neighborhood S_ε of $\mathbf{u}^*$ in $\mathcal{M}$ and a constant C_ε such that for every $\mathbf{u}_0 \in S_\varepsilon$

$$\|\mathbf{u}_k - \mathbf{u}^*\|_h \leq C_\varepsilon (r_\tau + \varepsilon)^k \|\mathbf{u}_0 - \mathbf{u}^*\|_h, \quad r_\tau := \max_i |1 - \tau \rho_i| = \max\{1 - \tau \rho_1, \tau \rho_{N-1} - 1\}. \quad (33)$$

The minimum of r_τ over τ is attained at $\tau^* = 2/(\rho_1 + \rho_{N-1})$ and equals $(\kappa(\mathbf{u}^*) - 1)/(\kappa(\mathbf{u}^*) + 1)$.

Proof The map F_τ is smooth near $\mathbf{u}^*$ because $\text{grad}E$ and the retraction are smooth, and $F_\tau(\mathbf{u}^*) = R_{\mathbf{u}^*}(0) = \mathbf{u}^*$. Let $\mathbf{u}(s)$ be a smooth curve on $\mathcal{M}$ with $\mathbf{u}(0) = \mathbf{u}^*$ and $\mathbf{u}'(0) = \xi$, and let $\mathbf{p}(s) := \text{grad}E(\mathbf{u}(s))$, viewed as a map into $\mathbb{R}^N$. Then $\mathbf{p}(0) = 0$ and $\mathbf{p}'(0) = D\text{grad}E(\mathbf{u}^*)[\xi]$. Since $\mathbf{p}(s) \in T_{\mathbf{u}(s)}\mathcal{M}$, the definition (16) gives

$$F_\tau(\mathbf{u}(s)) = \frac{\mathbf{u}(s) - \tau \mathbf{p}(s)}{d(s)}, \quad d(s) := \sqrt{1 + \tau^2 \|\mathbf{p}(s)\|_h^2},$$

with $d(0) = 1$ and $d'(0) = 0$ because $\mathbf{p}(0) = 0$. Differentiating at $s = 0$ gives $DF_\tau(\mathbf{u}^*)[\xi] = \xi - \tau \mathbf{p}'(0) = \xi - \tau D\text{grad}E(\mathbf{u}^*)[\xi]$. Differentiating $\langle \mathbf{p}(s), \mathbf{u}(s) \rangle_h = 0$ at $s = 0$ gives $\langle \mathbf{p}'(0), \mathbf{u}^* \rangle_h = 0$, so $D\text{grad}E(\mathbf{u}^*)[\xi]$ lies in $T_{\mathbf{u}^*}\mathcal{M}$. By Lemma B.1, $\text{Hess}E(\mathbf{u}^*)[\xi] = \nabla_\xi \text{grad}E = P_{\mathbf{u}^*}(D\text{grad}E(\mathbf{u}^*)[\xi] + \Gamma_{\mathbf{u}^*}(\xi, \text{grad}E(\mathbf{u}^*)))$. Since $\text{grad}E(\mathbf{u}^*) = 0$ and $D\text{grad}E(\mathbf{u}^*)[\xi]$ is tangent, this gives $\text{Hess}E(\mathbf{u}^*)[\xi] = D\text{grad}E(\mathbf{u}^*)[\xi]$, which is (32). This also follows from Boumal (2023, Prop. 5.3). Now let $\Phi := R_{\mathbf{u}^*}$, a smooth bijection from $T_{\mathbf{u}^*}\mathcal{M}$ onto the open hemisphere around $\mathbf{u}^*$ with smooth inverse (Section 2.3), and $\tilde{F} := \Phi^{-1} \circ F_\tau \circ \Phi$ near 0, so that $\tilde{F}(0) = 0$ and $D\tilde{F}(0) = DF_\tau(\mathbf{u}^*)$ because $D\Phi(0) = \mathbb{I}$. Since $\text{Hess}E(\mathbf{u}^*)$ is self-adjoint with respect to $g_{\mathbf{u}^*}$, the operator $D\tilde{F}(0) = \mathbb{I} - \tau \text{Hess}E(\mathbf{u}^*)$ is self-adjoint with eigenvalues $1 - \tau \rho_i$, and its operator norm in the norm $\|\cdot\|_{\mathbf{u}^*}$ on $T_{\mathbf{u}^*}\mathcal{M}$ is r_τ , which is smaller than one exactly when $0 < \tau < 2/\rho_{N-1}$. Let $0 < \varepsilon < 1 - r_\tau$, which suffices because the right side of (33) increases with ε . Since $\tilde{F}$ is smooth, there is a ball B_ε around 0 in the norm $\|\cdot\|_{\mathbf{u}^*}$, which we take inside $\{\|\xi\|_h \leq 1\}$, with the operator norm $\|D\tilde{F}(\xi)\|_{\mathbf{u}^*} \leq r_\tau + \varepsilon$ for $\xi \in B_\varepsilon$. The mean value inequality then gives $\|\tilde{F}(\xi)\|_{\mathbf{u}^*} \leq (r_\tau + \varepsilon)\|\xi\|_{\mathbf{u}^*}$ on B_ε , compare Henning (2023, Prop. 3.1) and Boumal (2023, Theorem 4.19). Let $\mathbf{u}_0$ lie in the neighborhood $S_\varepsilon := \Phi(B_\varepsilon)$ of $\mathbf{u}^*$, and let $\xi_0 := \Phi^{-1}(\mathbf{u}_0)$ and $\xi_{k+1} := \tilde{F}(\xi_k)$. Then every ξ_k lies in B_ε and $\|\xi_k\|_{\mathbf{u}^*} \leq (r_\tau + \varepsilon)^k \|\xi_0\|_{\mathbf{u}^*}$. By the definition of $\tilde{F}$,

$$\Phi(\xi_{k+1}) = F_\tau(\Phi(\xi_k)) = R_{\Phi(\xi_k)}(-\tau \text{grad}E(\Phi(\xi_k))),$$

and $\Phi(\xi_0) = \mathbf{u}_0$, so the points $\Phi(\xi_k)$ are exactly the iterates $\mathbf{u}_k$ of (16) with the constant step τ . The fixed chart Φ only supplies coordinates for the iterates, and the retraction in F_τ stays based at the current

iterate $\mathbf{u}_k$. By Lemma 4.1, $\|\xi_k\|_h \leq c_1^{-1/2} \|\xi_k\|_{\mathbf{u}^*} \leq c_1^{-1/2} (r_\tau + \varepsilon)^k \|\xi_0\|_{\mathbf{u}^*} \leq \sqrt{c_2/c_1} (r_\tau + \varepsilon)^k \|\xi_0\|_h$. For $\mathbf{u} \in \mathcal{M}$ and $\xi \in T_{\mathbf{u}}\mathcal{M}$, the definition (16) and $\langle \mathbf{u}, \xi \rangle_h = 0$ give

$$\|R_{\mathbf{u}}(\xi) - \mathbf{u}\|_h^2 = 2(1 - (1 + \|\xi\|_h^2)^{-1/2}) = \|\xi\|_h^2 \frac{2}{\sqrt{1 + \|\xi\|_h^2} (1 + \sqrt{1 + \|\xi\|_h^2})}, \quad (34)$$

where the last factor decreases from 1 at $\xi = 0$ to $2 - \sqrt{2}$ at $\|\xi\|_h = 1$. Hence $\|R_{\mathbf{u}}(\xi) - \mathbf{u}\|_h \leq \|\xi\|_h$ for every $\xi \in T_{\mathbf{u}}\mathcal{M}$, and $\|R_{\mathbf{u}}(\xi) - \mathbf{u}\|_h \geq \|\xi\|_h/2$ for $\|\xi\|_h \leq 1$. With $\mathbf{u} = \mathbf{u}^*$, the first bound gives $\|\mathbf{u}_k - \mathbf{u}^*\|_h \leq \|\xi_k\|_h$, and the second gives $\|\xi_0\|_h \leq 2\|\mathbf{u}_0 - \mathbf{u}^*\|_h$ because $\xi_0 \in \mathcal{B}_\varepsilon$. Therefore

$$\|\mathbf{u}_k - \mathbf{u}^*\|_h \leq \|\xi_k\|_h \leq \sqrt{c_2/c_1} (r_\tau + \varepsilon)^k \|\xi_0\|_h \leq 2\sqrt{c_2/c_1} (r_\tau + \varepsilon)^k \|\mathbf{u}_0 - \mathbf{u}^*\|_h,$$

which is (33) with $C_\varepsilon := 2\sqrt{c_2/c_1}$. Finally, r_τ is the maximum of the decreasing function $1 - \tau\rho_1$ and the increasing function $\tau\rho_{N-1} - 1$, so it is minimized where the two are equal, at $\tau^* = 2/(\rho_1 + \rho_{N-1})$, and its minimum is $1 - \tau^*\rho_1 = (\rho_{N-1} - \rho_1)/(\rho_{N-1} + \rho_1) = (\kappa(\mathbf{u}^*) - 1)/(\kappa(\mathbf{u}^*) + 1)$. $\square$

For the a_u metric, (19) reads

$$\rho_{\mathbf{u}^*}(\eta) = 1 - \frac{\langle (\lambda_1(A_\beta)\mathbb{I} - 2\beta \text{diag}((\mathbf{u}^*)^2))\eta, \eta \rangle_h}{\langle A_\beta \eta, \eta \rangle_h}.$$

Up to the space discretization, the eigenvalues $1 - \rho_i$ of $DF_1(\mathbf{u}^*)$ in (32) are thus the eigenvalues of the bilinear form $(v, (\lambda - 2\beta|u|^2)w)$ relative to the inner product $(v, w)_{a_u}$ on the L^2 -orthogonal complement of u . This is the weighted eigenvalue problem in Henning (2023). The undamped inverse iteration $\mathbf{u}_{k+1} \propto A_\beta(\mathbf{u}_k)^{-1}\mathbf{u}_k$ is the iteration (16) under the a_u metric with $\tau_k = 1$, that is, $\mathbf{u}_{k+1} = F_1(\mathbf{u}_k)$. Under Assumption 2.1, $\rho_{N-1} < 2$ by Corollary 3.5, so Theorem 4.2 applies with $\tau = 1$, and this iteration converges locally with the rate $r_1 + \varepsilon$ for every $\varepsilon > 0$, where $r_1 = \max_i |1 - \rho_i|$ is the largest modulus of the eigenvalues of the weighted problem, in agreement with Henning (2023, Remark 3.3). For a fixed step $\tau \in (0, 2/\rho_{N-1})$ the eigenvalues of $DF_\tau(\mathbf{u}^*)$ are $1 - \tau\rho_i = 1 - \tau + \tau\mu_i$ with $\mu_i := 1 - \rho_i$. Thus the rate $r_\tau = \max_i |1 - \tau + \tau\mu_i|$ of Theorem 4.2 is the sharper rate noted in Henning (2023, Section 6.1) for the damped inverse iteration with damping parameter τ . By Corollary 3.5, $r_\tau \leq |1 - \tau| + \tau \max_i |\mu_i| \leq |1 - \tau| + \tau\lambda_1(A_\beta)/\lambda_2(A_\beta)$, which is the bound of Henning (2023, Theorem 6.2). The critical damping parameter $2(1 + \lambda_1(A_\beta)/\lambda_2(A_\beta))^{-1}$ of that bound is at most $2/\rho_{N-1}$, also by Corollary 3.5.

The same spectrum also controls the energy on a whole sublevel set. The proof combines sufficient decrease with a Polyak–Łojasiewicz inequality, as in Boumal (2023, Lemma 11.28 and Theorem 11.29) for geodesically strongly convex functions, with geodesic convexity replaced by bounds on the second derivative along retraction curves. Absil et al. (2008, Theorem 4.5.6) prove an asymptotic energy contraction for Armijo steps along $-\text{grad}E$, with a rate that also depends on the ratio of the extreme eigenvalues of the Hessian.

PROPOSITION 4.3 (Energy contraction on a sublevel set) Let Assumption 2.1 hold and let $\rho_- := \rho_1$, $\rho_+ := \rho_{N-1}$. There is $\varepsilon > 0$ such that, with $S_0 := \{\mathbf{u} \in \mathcal{M} : E(\mathbf{u}) \leq E(\mathbf{u}^*) + \varepsilon, \langle \mathbf{u}, \mathbf{u}^* \rangle_h > 0\}$, the

iteration

$$\mathbf{u}_{k+1} = R_{\mathbf{u}_k} \left(-\frac{1}{2\rho_+} \text{grad} E(\mathbf{u}_k) \right)$$

started from any $\mathbf{u}_0 \in S_0$ stays in S_0 and satisfies

$$E(\mathbf{u}_k) - E(\mathbf{u}^*) \leq \left(1 - \frac{1}{4\kappa(\mathbf{u}^*)} \right)^k (E(\mathbf{u}_0) - E(\mathbf{u}^*)). \quad (35)$$

Proof Step 1. The function $Q(t, \mathbf{u}, \eta) := \frac{d^2}{dt^2} E(R_{\mathbf{u}}(t\eta))$ is continuous on $[-1, 1] \times K$, and $Q(0, \mathbf{u}^*, \eta) = g_{\mathbf{u}^*}(\text{Hess} E(\mathbf{u}^*)[\eta], \eta) \in [\rho_-, \rho_+]$ for $(\mathbf{u}^*, \eta) \in K$ by (18). There are $r \in (0, 1]$ and a neighborhood $\mathcal{N}$ of $\mathbf{u}^*$ with $\rho_-/2 \leq Q \leq 2\rho_+$ for $|t| \leq r$, $\mathbf{u} \in \mathcal{N}$, $(\mathbf{u}, \eta) \in K$. Otherwise there would be $t_j \rightarrow 0$ and $(\mathbf{u}_j, \eta_j) \in K$ with $\mathbf{u}_j \rightarrow \mathbf{u}^*$ and $Q(t_j, \mathbf{u}_j, \eta_j) \notin (\rho_-/2, 2\rho_+)$. Since K is compact, a subsequence of $(\mathbf{u}_j, \eta_j)$ converges to some $(\mathbf{u}^*, \eta^*) \in K$, and continuity of Q gives $Q(0, \mathbf{u}^*, \eta^*) \notin (\rho_-/2, 2\rho_+)$, which contradicts $Q(0, \mathbf{u}^*, \eta^*) \in [\rho_-, \rho_+]$. Now let $\mathbf{u} \in \mathcal{N}$ and $\mathbf{w} \in T_{\mathbf{u}}\mathcal{M}$ with $0 < \|\mathbf{w}\|_{\mathbf{u}} \leq r$, and set $\eta := \mathbf{w}/\|\mathbf{w}\|_{\mathbf{u}}$. Since $\frac{d}{dt} E(R_{\mathbf{u}}(t\eta))|_{t=0} = DE(\mathbf{u})[\eta] = g_{\mathbf{u}}(\text{grad} E(\mathbf{u}), \eta)$ by (16) and (15), Taylor's theorem for $t \mapsto E(R_{\mathbf{u}}(t\eta))$ at $t = \|\mathbf{w}\|_{\mathbf{u}}$ gives

$$\frac{\rho_-}{4} \|\mathbf{w}\|_{\mathbf{u}}^2 \leq E(R_{\mathbf{u}}(\mathbf{w})) - E(\mathbf{u}) - g_{\mathbf{u}}(\text{grad} E(\mathbf{u}), \mathbf{w}) \leq \rho_+ \|\mathbf{w}\|_{\mathbf{u}}^2. \quad (36)$$

Step 2. Since $E(-\mathbf{u}) = E(\mathbf{u})$ and $\pm \mathbf{u}^*$ are the only global minimizers (Lemma 2.2), compactness of $\mathcal{M}$ shows that for small ε the sublevel set $\{E \leq E(\mathbf{u}^*) + \varepsilon\}$ lies in two disjoint balls of any prescribed radius around $\pm \mathbf{u}^*$, and S_0 is its part in the ball around $\mathbf{u}^*$. For $\mathbf{u} \in S_0$ let $\xi_{\mathbf{u}} := \mathbf{u}^* / \langle \mathbf{u}, \mathbf{u}^* \rangle_h - \mathbf{u} \in T_{\mathbf{u}}\mathcal{M}$ be the tangent vector with $R_{\mathbf{u}}(\xi_{\mathbf{u}}) = \mathbf{u}^*$ (Section 2.3). Since $\|\xi_{\mathbf{u}}\|_h^2 = \langle \mathbf{u}, \mathbf{u}^* \rangle_h^{-2} - 1$, it satisfies $\|\xi_{\mathbf{u}}\|_{\mathbf{u}}^2 \leq c_2(\langle \mathbf{u}, \mathbf{u}^* \rangle_h^{-2} - 1)$. We take ε so small that $S_0 \subset \mathcal{N}$ and that $\|\text{grad} E(\mathbf{u})\|_{\mathbf{u}}/(2\rho_+) < r$ and $\|\xi_{\mathbf{u}}\|_{\mathbf{u}} < r$ on S_0 , which is possible because both quantities are continuous and vanish at $\mathbf{u}^*$. *Step 3.* Let $\mathbf{u} \in S_0$ with $\mathbf{g} := \text{grad} E(\mathbf{u}) \neq 0$. For $t \in (0, 1]$ the vector $\mathbf{w}_t := -t\mathbf{g}/(2\rho_+)$ satisfies $\|\mathbf{w}_t\|_{\mathbf{u}} < r$ and $g_{\mathbf{u}}(\mathbf{g}, \mathbf{w}_t) = -t\|\mathbf{g}\|_{\mathbf{u}}^2/(2\rho_+)$, so the upper bound in (36) gives $E(R_{\mathbf{u}}(\mathbf{w}_t)) \leq E(\mathbf{u}) - t(2-t)\|\mathbf{g}\|_{\mathbf{u}}^2/(4\rho_+) \leq E(\mathbf{u})$. Hence the retraction segment from $\mathbf{u}$ to $\mathbf{u}_+ := R_{\mathbf{u}}(\mathbf{w}_1)$ stays in the sublevel set. Since the segment is connected and starts at $\mathbf{u} \in S_0$, and the sublevel set lies in two disjoint balls, the segment stays in S_0 , and $t = 1$ gives $E(\mathbf{u}_+) \leq E(\mathbf{u}) - \|\mathbf{g}\|_{\mathbf{u}}^2/(4\rho_+)$. The lower bound in (36) with $\mathbf{w} = \xi_{\mathbf{u}}$ gives

$$E(\mathbf{u}^*) \geq E(\mathbf{u}) + g_{\mathbf{u}}(\mathbf{g}, \xi_{\mathbf{u}}) + \frac{\rho_-}{4} \|\xi_{\mathbf{u}}\|_{\mathbf{u}}^2 \geq E(\mathbf{u}) - \|\mathbf{g}\|_{\mathbf{u}}^2/\rho_-,$$

because $g_{\mathbf{u}}(\mathbf{g}, \xi) + \frac{\rho_-}{4} \|\xi\|_{\mathbf{u}}^2 \geq -\|\mathbf{g}\|_{\mathbf{u}}^2/\rho_-$ for every ξ , with equality at $\xi = -2\mathbf{g}/\rho_-$. Hence $\|\mathbf{g}\|_{\mathbf{u}}^2 \geq \rho_-(E(\mathbf{u}) - E(\mathbf{u}^*))$, and the two bounds give $E(\mathbf{u}_+) - E(\mathbf{u}^*) \leq (1 - \rho_-/(4\rho_+))(E(\mathbf{u}) - E(\mathbf{u}^*))$. If $\text{grad} E(\mathbf{u}) = 0$ for some $\mathbf{u} \in S_0$, then $\mathbf{u} = \mathbf{u}^*$, because otherwise the same lower bound gives $E(\mathbf{u}^*) \geq E(\mathbf{u}) + \frac{\rho_-}{4} \|\xi_{\mathbf{u}}\|_{\mathbf{u}}^2 > E(\mathbf{u})$. The iteration then stays at $\mathbf{u}^*$, and (35) follows by induction. $\square$

4.2. Global convergence

THEOREM 4.4 (Global convergence to a single critical point) Let $(\mathbf{u}_k)$ be generated by (16) from any $\mathbf{u}_0 \in \mathcal{M}$ with the constant step $\tau_k = 1/C_E$, and write $\mathbf{g}_k := \text{grad} E(\mathbf{u}_k)$. Then

$$E(\mathbf{u}_{k+1}) \leq E(\mathbf{u}_k) - \frac{1}{2C_E} \|\mathbf{g}_k\|_{\mathbf{u}_k}^2. \quad (37)$$

In addition, $\text{grad} E(\mathbf{u}_k) \rightarrow 0$ and the whole sequence converges to a single critical point $\mathbf{u}_{\infty} \in \mathcal{M}$ with $\sum_k \|\mathbf{u}_{k+1} - \mathbf{u}_k\|_h < \infty$. The same holds, with adjusted constants, for every step size rule satisfying (37)

and $\|\mathbf{u}_{k+1} - \mathbf{u}_k\|_h \leq C \|\mathbf{g}_k\|_{\mathbf{u}_k}$, in particular for the exact line search $\tau_k = \arg \min_{\tau \in [0, \tau_{\max}]} E(R_{\mathbf{u}_k}(-\tau \mathbf{g}_k))$ with $\tau_{\max} \geq 1/C_E$.

Proof Step 1. If $\mathbf{g}_k = 0$, then $\mathbf{u}_{k+1} = \mathbf{u}_k$ and (37) is trivial. Otherwise, with $\boldsymbol{\eta} := -\mathbf{g}_k / \|\mathbf{g}_k\|_{\mathbf{u}_k}$, $s_k := \|\mathbf{g}_k\|_{\mathbf{u}_k} / C_E \leq C_g$ and $\hat{E}(t) := E(R_{\mathbf{u}_k}(t\boldsymbol{\eta}))$, we have $\hat{E}'(0) = g_{\mathbf{u}_k}(\mathbf{g}_k, \boldsymbol{\eta}) = -\|\mathbf{g}_k\|_{\mathbf{u}_k}$ as in Step 1 of the proof of Proposition 4.3, and $|\hat{E}''| \leq C_E$ on $[0, s_k]$ by the definition of C_E , because $(\mathbf{u}_k, \boldsymbol{\eta}) \in K$ and $s_k \leq C_g$. Taylor's theorem gives $E(\mathbf{u}_{k+1}) \leq E(\mathbf{u}_k) - s_k \|\mathbf{g}_k\|_{\mathbf{u}_k} + \frac{C_E}{2} s_k^2$, which is (37), and the exact line search on $[0, \tau_{\max}]$ does at least as well as $\tau = 1/C_E$. By (34), $\|R_{\mathbf{u}}(\xi) - \mathbf{u}\|_h \leq \|\xi\|_h$ for $\xi \in T_{\mathbf{u}}\mathcal{M}$. Since $\|\xi\|_h \leq \|\xi\|_{\mathbf{u}} / \sqrt{c_1}$, the step $\xi = -\tau_k \mathbf{g}_k$ gives the step bound $\|\mathbf{u}_{k+1} - \mathbf{u}_k\|_h \leq C \|\mathbf{g}_k\|_{\mathbf{u}_k}$ with $C = 1/(\sqrt{c_1} C_E)$ for $\tau_k = 1/C_E$ and $C = \tau_{\max} / \sqrt{c_1}$ for the exact line search. For a general rule, the step bound is assumed. Summing (37) over k and using $E \geq 0$ gives $\sum_k \|\mathbf{g}_k\|_{\mathbf{u}_k}^2 \leq 2C_E E(\mathbf{u}_0) < \infty$, so $\|\mathbf{g}_k\|_h \leq c_1^{-1/2} \|\mathbf{g}_k\|_{\mathbf{u}_k} \rightarrow 0$ by Lemma 4.1, and the energies decrease to a limit E_∞ . Since $\mathcal{M}$ is compact and $\text{grad} E$ is continuous, the set $\mathcal{A}$ of accumulation points of $(\mathbf{u}_k)$ is nonempty and compact, and every $\mathbf{v} \in \mathcal{A}$ is a critical point with $E(\mathbf{v}) = E_\infty$. For every metric except the Thomas–Fermi metric, the criticality of accumulation points and the convergence of the gradient to zero also follow from Absil et al. (2008, Theorem 4.3.1 and Corollary 4.3.2) and Boumal (2023, Prop. 4.7), since both step rules achieve at least the decrease of the Armijo trial step $1/C_E$. For the modified H^1 metric, this sufficient-decrease argument parallels (Chen et al., 2025, Section 5.1).

Step 2. Let $\mathbf{v} \in \mathcal{M}$. The retraction $R_{\mathbf{v}}$ is a real-analytic parametrization of $\mathcal{M}$ around $\mathbf{v}$ by (16) and Section 2.3, and E is a polynomial. Hence the Łojasiewicz inequality (Łojasiewicz, 1963) holds for the projected gradient (Schneider and Uschmajew, 2015, Prop. 2.2): there are $C_{\mathbf{v}} > 0$, $\varepsilon_{\mathbf{v}} > 0$ and $\theta_{\mathbf{v}} \in (0, \frac{1}{2}]$ with

$$|E(\mathbf{u}) - E(\mathbf{v})|^{1-\theta_{\mathbf{v}}} \leq C_{\mathbf{v}} \|\mathbf{q}(\mathbf{u})\|_h \quad \text{for } \mathbf{u} \in \mathcal{M}, \|\mathbf{u} - \mathbf{v}\|_h < \varepsilon_{\mathbf{v}},$$

where $\mathbf{q}(\mathbf{u}) := A_{\beta}(\mathbf{u})\mathbf{u} - \lambda(\mathbf{u})\mathbf{u}$ is the gradient (15) for the metric $B = \mathbb{I}$, that is, the projection of $A_{\beta}(\mathbf{u})\mathbf{u}$ onto $T_{\mathbf{u}}\mathcal{M}$ in $\langle \cdot, \cdot \rangle_h$. The change of variables $\mathbf{x} = \mathbb{M}^{1/2}\mathbf{u}$ identifies $\langle \cdot, \cdot \rangle_h$ with the Euclidean inner product, and Schneider and Uschmajew (2015, Prop. 2.2) is applied in these coordinates, where the projected Euclidean gradient of E is $\mathbb{M}^{1/2}\mathbf{q}(\mathbf{u})$ with norm $\|\mathbf{q}(\mathbf{u})\|_h$. For $\zeta \in T_{\mathbf{u}}\mathcal{M}$, Lemma 4.1 gives $\langle \mathbf{q}(\mathbf{u}), \zeta \rangle_h = DE(\mathbf{u})[\zeta] = g_{\mathbf{u}}(\text{grad} E(\mathbf{u}), \zeta) \leq \|\text{grad} E(\mathbf{u})\|_{\mathbf{u}} \|\zeta\|_{\mathbf{u}} \leq \sqrt{c_2} \|\text{grad} E(\mathbf{u})\|_{\mathbf{u}} \|\zeta\|_h$, and $\zeta = \mathbf{q}(\mathbf{u})$ gives $\|\mathbf{q}(\mathbf{u})\|_h \leq \sqrt{c_2} \|\text{grad} E(\mathbf{u})\|_{\mathbf{u}}$. Hence

$$|E(\mathbf{u}) - E(\mathbf{v})|^{1-\theta_{\mathbf{v}}} \leq \sqrt{c_2} C_{\mathbf{v}} \|\text{grad} E(\mathbf{u})\|_{\mathbf{u}} \quad \text{for } \mathbf{u} \in \mathcal{M}, \|\mathbf{u} - \mathbf{v}\|_h < \varepsilon_{\mathbf{v}}. \quad (38)$$

Now let $\mathbf{v}_1, \dots, \mathbf{v}_m \in \mathcal{A}$ be such that the balls $\{\mathbf{u} \in \mathcal{M} : \|\mathbf{u} - \mathbf{v}_i\|_h < \varepsilon_{\mathbf{v}_i}\}$ cover the compact set $\mathcal{A}$, let $\mathcal{U}$ be their union, $\theta := \min_i \theta_{\mathbf{v}_i}$ and $C_L := \sqrt{c_2} \max_i C_{\mathbf{v}_i}$. Since $E(\mathbf{v}_i) = E_\infty$ and $t^{1-\theta} \leq t^{1-\theta_{\mathbf{v}_i}}$ for $0 \leq t \leq 1$, (38) gives

$$|E(\mathbf{u}) - E_\infty|^{1-\theta} \leq C_L \|\text{grad} E(\mathbf{u})\|_{\mathbf{u}} \quad \text{for } \mathbf{u} \in \mathcal{U} \text{ with } |E(\mathbf{u}) - E_\infty| \leq 1. \quad (39)$$

All but finitely many $\mathbf{u}_k$ lie in $\mathcal{U}$, because otherwise a subsequence in the closed set $\mathcal{M} \setminus \mathcal{U}$ would have an accumulation point outside $\mathcal{U}$, which contradicts $\mathcal{A} \subset \mathcal{U}$.

Step 3. If some $\mathbf{g}_{k_0} = 0$ the sequence is eventually constant. Otherwise $e_k := E(\mathbf{u}_k) - E_\infty > 0$ decreases to zero, and there is k_0 such that $\mathbf{u}_k \in \mathcal{U}$ and $e_k \leq 1$ for all $k \geq k_0$. For $k \geq k_0$, concavity

of $r \mapsto r^\theta$ gives the first inequality, (37) and (39) give the second and the step bound gives the last inequality in

$$\begin{aligned} e_k^\theta - e_{k+1}^\theta &\geq \theta e_k^{\theta-1} (e_k - e_{k+1}) \geq \frac{\theta}{C_L \|\mathbf{g}_k\|_{\mathbf{u}_k}} \cdot \frac{\|\mathbf{g}_k\|_{\mathbf{u}_k}^2}{2C_E} \\ &= \frac{\theta}{2C_L C_E} \|\mathbf{g}_k\|_{\mathbf{u}_k} \geq c_\star \|\mathbf{u}_{k+1} - \mathbf{u}_k\|_h, \end{aligned}$$

where $c_\star := \theta / (2C_L C_E C)$. Summing over $k \geq k_0$ gives $\sum_{k \geq k_0} \|\mathbf{u}_{k+1} - \mathbf{u}_k\|_h \leq c_\star^{-1} e_{k_0}^\theta < \infty$. Hence $(\mathbf{u}_k)$ is Cauchy and converges to a point $\mathbf{u}_\infty$, which is then the only accumulation point, so $\mathbf{u}_\infty \in \mathcal{A}$ is a critical point with $E(\mathbf{u}_\infty) = E_\infty$. $\square$

5. Numerical experiments

This section tests the predictions of Section 3 in three dimensions with up to 10^9 unknowns, where the cost of applying each metric enters the comparison. Each metric is tested with Riemannian gradient descent, which Section 4 analyzes, and with the Riemannian conjugate gradient method. The convergence analysis does not cover conjugate gradients. However, the practical performance of conjugate gradient methods is related to the conditioning of the Hessian (Zheng et al., 2025), which Section 3 analyzes. The tests below show that both methods behave under the six metrics as the Hessian analysis suggests. All tests were run in Python with JAX on one NVIDIA GH200 GPU card, in double precision unless stated otherwise.

5.1. Setup of the three-dimensional comparisons

The domain is $\Omega = (-8, 8)^3$ with $\beta = 1600$ and the stirrer potential

$$V(\mathbf{x}) = x_1^2 + x_2^2 + 4x_3^2 + 2w_0 e^{-(x_1 - r_0)^2 - x_2^2}, \quad w_0 = 30, \quad r_0 = 1, \quad (40)$$

a standard model of a laser stirrer in a condensate (Bao and Du, 2004). Its harmonic part is separable and is taken as $\mathbb{V}_1$, and the Gaussian beam, which couples x_1 and x_2 , is $\mathbb{V}_2$. The discretization is the $\mathcal{Q}^1$ element with Gauss–Lobatto quadrature on a uniform mesh, that is, the second-order finite difference scheme, with $n = 499$ interior points per direction, so $N = 499^3 \approx 1.2 \times 10^8$. Every method starts from the Thomas–Fermi profile $\mathbf{u}_0 \propto \sqrt{(\lambda_{\text{TF}} - V)_+ / \beta}$, with λ_{TF} defined by $\int_\Omega (\lambda_{\text{TF}} - V)_+ \, d\mathbf{x} = \beta$. The shift in the modified H^1 metric is $\alpha = 20$. At $\beta = 1600$ the modified H^1 flow takes 582, 207, 169 and 182 iterations at $\alpha = 1, 5, 20$ and 100, and its conjugate gradient variant takes 62, 37, 31 and 32. At $\alpha = 0$, which gives the classical H^1 seminorm, gradient descent does not reach 10^{-12} in 600 iterations and conjugate gradients needs 114. Among these values, $\alpha = 20$ gives the fewest iterations for both methods. Conjugate gradients is much less sensitive to α than gradient descent.

The six metrics of Table 1 are compared. The Thomas–Fermi metric uses $\ell = 1/2$, and $\lambda(\mathbf{u})$ in (12) is recomputed from the energy evaluation without a solve. The operators $-\Delta_h + \alpha \mathbb{I}$ and $-\Delta_h + \mathbb{V}_1$ are separable for the potential (40) and are applied exactly by tensor-product direct solvers (Liu et al., 2024; Liu and Zhang, 2026). The other four are applied by conjugate gradients preconditioned with $(-\Delta_h + \mathbb{V}_1)^{-1}$, stopped at relative residual 10^{-8} .

Every method takes the step size that minimizes the energy along the retracted ray of its search direction η_k , $\tau_k = \arg \min_\tau E(R_{\mathbf{u}_k}(\tau \eta_k))$, where $\eta_k = -\text{grad} E(\mathbf{u}_k)$ for gradient descent. This costs two

applications of $-\Delta_h + \mathbb{V}$ and no solve, since the energy along the ray is a rational function of τ whose coefficients are inner products of $\mathbf{u}_k$ and η_k . Conjugate gradients is the geometric conjugate gradient method (Absil et al., 2008, Algorithm 13) with the Fletcher–Reeves parameter (Absil et al., 2008, (8.28)), in which each gradient norm is evaluated in the metric at its own iterate. The method transports the previous direction by the tangent projection (15), as in Absil et al. (2008, (8.10)), and restarts to steepest descent whenever the result is not a descent direction. Besides the wall clock time, we report the machine-independent number of applications of the tensor-product solver. An exact metric application counts as one, and an inexact one counts one per inner conjugate gradient iteration plus two for the separable warm start and the initial residual of each inner solve. All runs stop at $|E(\mathbf{u}_k) - E^*| < 10^{-12}$, where the reference $E^* = 9.012256553261389$ is computed on the same grid by a Davis–Yin splitting scheme (Li et al., 2026), which is not part of the comparison and is run until its residual stops decreasing. As a check of the discretization, Richardson extrapolation from the 499^3 and 799^3 grids reproduces the Q^{20} value of Section 5.2 to 6×10^{-10} . On the larger boxes of Sections 5.3 and 5.4, the step size of the Davis–Yin scheme decreases like L^{-2} , so the references for $\beta = 6400$ and 102400 and for $L = 32$ are computed by Riemannian conjugate gradients in the modified a_0 metric, run until the residual stops decreasing.

5.2. Comparison of the metrics

Figure 1 reports the comparison on the 499^3 grid. Among the three classical metrics, the a_u metric needs the fewest iterations but is not the fastest, because its operator depends on the iterate and is applied inexactly, with about 17 preconditioner applications per iteration. The modified a_0 metric needs about as many iterations as the a_0 metric (Proposition 3.6). Since each application of this metric is a single tensor-product solve, it needs 244 applications of the solver against 3994 for the a_0 metric and is an order of magnitude faster. Its conjugate gradient variant and that of the modified H^1 metric are the fastest methods. Conjugate gradients reduces the iteration count most where gradient descent is slowest, as expected if the rate of conjugate gradients is governed by $\sqrt{\kappa}$ rather than by κ .

The modified a_u metric needs fewer iterations than the a_u metric (Proposition 3.8). The halving of the condition number is asymptotic in β and is not yet reached at $\beta = 1600$. Each iteration applies the metric twice, however, and $(-\Delta_h + \mathbb{V}_1)^{-1}$ preconditioners $-\Delta_h + \mathbb{V} + 3\beta \text{diag}(\mathbf{u}^2)$ less well than $A_\beta(\mathbf{u})$, so both variants are about twice as slow in wall clock time.

The Thomas–Fermi metric needs the fewest iterations. Although it needs two solves per iteration, it is faster than the a_u flow, and among the gradient flows with an inexactly applied metric it uses the fewest solver applications. Its inner solves are the most expensive, 36.6 preconditioner applications per metric application against 17.0 for the a_u metric and 23.7 for the modified a_u metric, because $-\Delta_h + \mathbb{V}_1$ is a poor approximation of an operator whose potential is clipped at $\lambda(\mathbf{u})$. At $\beta = 1600$ the small number of outer iterations compensates for this cost.

The iteration counts are essentially mesh independent. The five metrics other than the Thomas–Fermi metric were also run on a 99^3 grid, with 125 times fewer unknowns. On that grid, every iteration count differs from its value on the 499^3 grid by at most one.

The exact spectrum on the stirrer (40) at $\beta = 1600$, computed on the 499^3 grid as in Appendix A, accounts for the iteration counts of Figure 1. The condition numbers are 1.30 for the Thomas–Fermi metric, 4.73 for the modified a_u metric, 6.75 for the a_u metric, 20.8 for the a_0 metric, 24.3 for the modified a_0 metric and 39.9 for the modified H^1 metric. If the rate $(\kappa(\mathbf{u}^*) - 1)/(\kappa(\mathbf{u}^*) + 1)$ of Theorem 4.2 governs the number of iterations k to a fixed tolerance, $2k \log((\kappa(\mathbf{u}^*) + 1)/(\kappa(\mathbf{u}^*) - 1))$ is the same for every metric. It lies between 20.1 and 21.5 for the four metrics with $4.7 \leq \kappa(\mathbf{u}^*) \leq 24.3$,

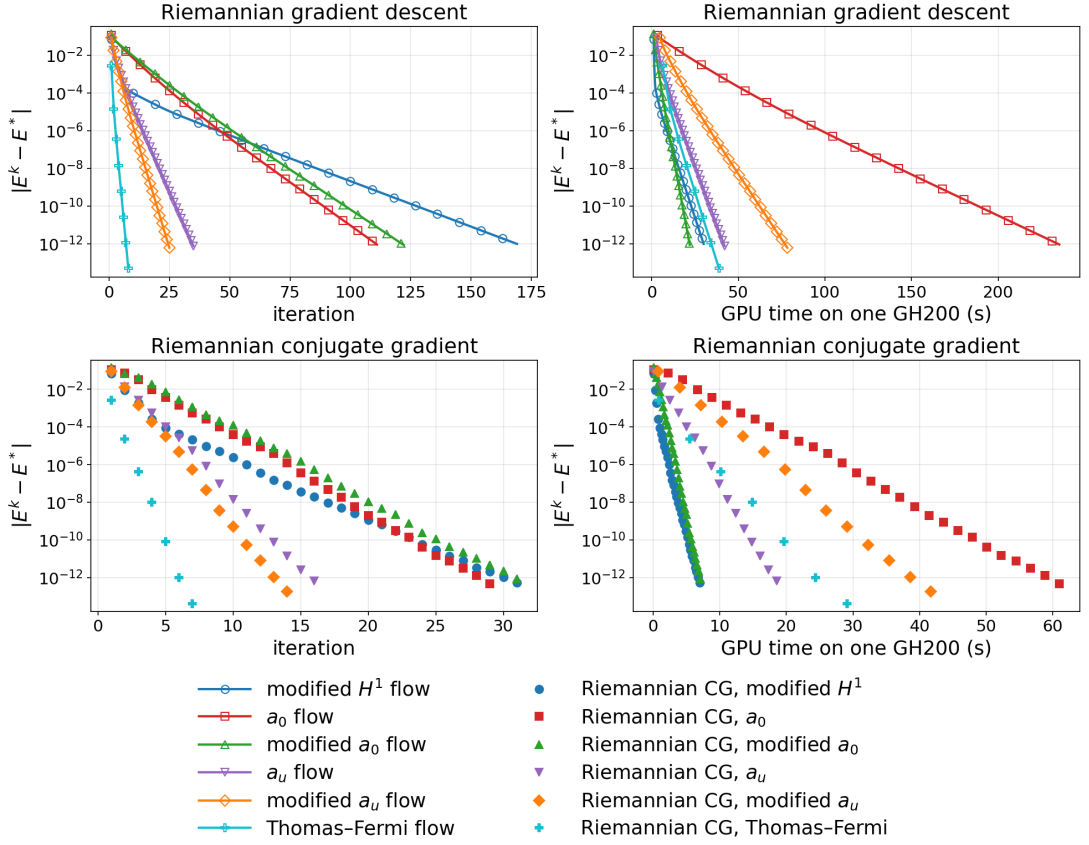


FIG. 1. The metrics on the stirrer potential (40) at $\beta = 1600$ from the Thomas–Fermi start, on a 499^3 grid, with $\ell = 1/2$ in the Thomas–Fermi metric: the energy error against the iteration count on the left and against GPU time on the right, gradient descent in the top row and conjugate gradients in the bottom one. Color and marker indicate the metric, and the horizontal scales differ between the rows.

whose iteration counts differ by a factor of five, and between 25.5 and 27.8 for their conjugate gradient variants with $\sqrt{\kappa(\mathbf{u}^*)}$ in place of $\kappa(\mathbf{u}^*)$. The other two metrics deviate in opposite directions. The maximizing eigenvectors, computed on the 199^3 grid, lie at the center of the trap for four metrics. For the modified a_u metric the maximizing eigenvector is the grid-scale mode that attains the upper edge of (23) where $\mathbb{V}$ is largest, and for the modified H^1 metric it also lies where $\mathbb{V}$ is largest and the iterates are negligibly small, so the metric needs fewer iterations than its condition number predicts. The Thomas–Fermi metric needs more, since its few iterations are too few for the asymptotic regime to be reached.

To test the comparison on a second discretization, the same problem is solved with Q^{20} spectral elements on 50 cells per direction, so $n = 999$ and the problem has about 10^9 unknowns. The mass matrix is then the diagonal of the Gauss–Lobatto weights rather than $h^3\mathbb{I}$, and the one-dimensional operators are diagonalized numerically. The reference $E^* = 9.012310296639$ is computed on the same grid by the Davis–Yin scheme, run until its residual stops decreasing, and the runs stop at $|E - E^*| <$

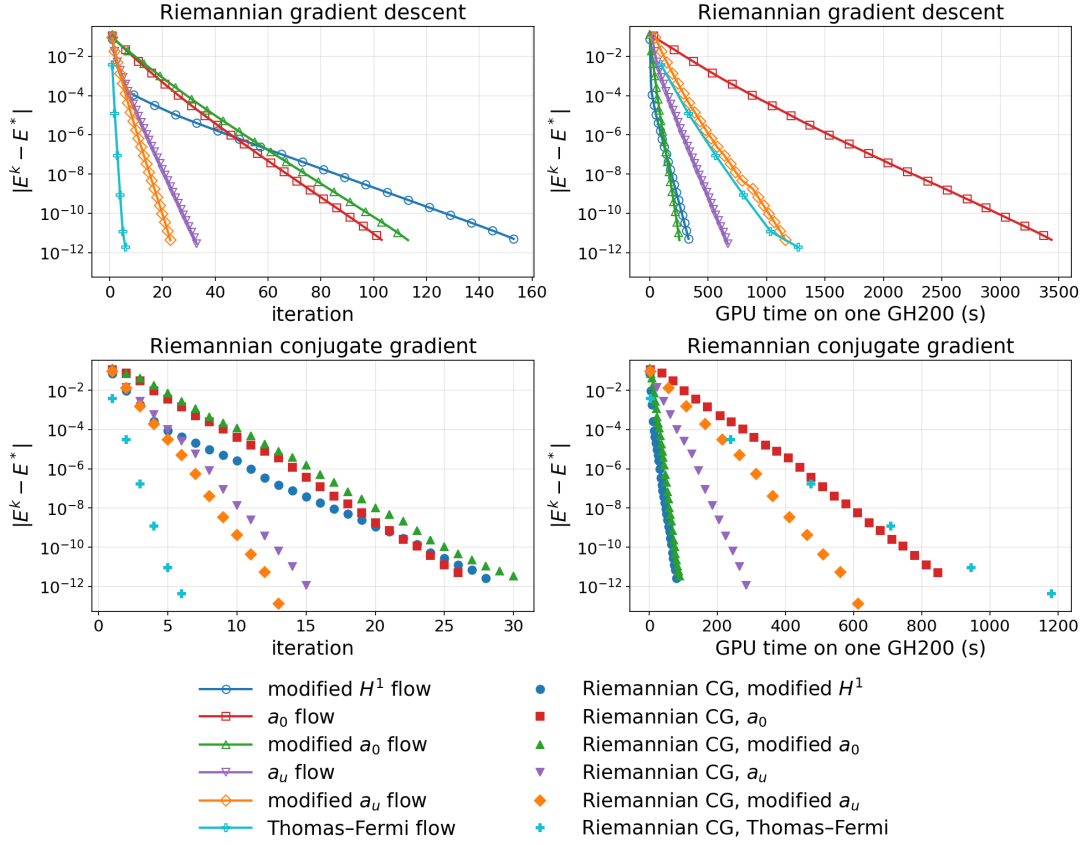


FIG. 2. The same comparison with Q^{20} spectral elements on a 999^3 grid, about 10^9 unknowns, in the panels and the drawing convention of Figure 1.

5×10^{-12} . Figure 2 shows that the iteration counts carry over from the 499^3 grid and are slightly lower throughout. The ordering by iterations is unchanged, and the two metrics with separable operators remain the cheapest per iteration. In wall clock time, however, the Thomas-Fermi flow is now slower than the a_u flow, whereas at 499^3 it was faster, and the difference comes from the inner solve.

In double precision one vector of 999^3 unknowns is 8 gigabytes, so the metrics that invert a nonseparable operator fit into memory only if the Krylov vectors of the inner solve are held in single precision, while the residual is formed and the correction is accumulated in double precision, as in [Liu and Zhang \(2026\)](#). At $n = 499$, where the double-precision solve also fits, the a_0 flow with the mixed-precision solve reaches the tolerance in the same number of iterations as with the double-precision one. Every method fits, with peak memory between 59.5 and 78.1 gigabytes of the 95.6 available, once $B(\mathbf{u})\mathbf{x}$ is consumed by a reduction and never stored. The Thomas-Fermi metric requires a further change of the inner solve. Its operator is the one that the separable preconditioner approximates worst, with about 75 inner iterations against 29 for $A_\beta(\mathbf{u})$. Over that many iterations, the recursively updated residual of the single-precision recurrence drifts so far from the true residual that the metric is applied far less accurately than the reported 10^{-6} , and the outer iteration stalls with a collapsing step size. Capping the

single-precision solve at 25 iterations and recomputing the residual in double precision six times, at the same 10^{-6} target, restores convergence. The two Thomas–Fermi runs of Figure 2 were computed with the capped solves, at about 2.4 times the inner work per iteration.

5.3. Dependence on the interaction strength

For the one-dimensional harmonic trap, Remark 3.12 shows that the condition number of the Thomas–Fermi metric is bounded in β , while that of the a_u metric grows like $\lambda_1(A_\beta)^{2/3}$, those of its family (24) at least as fast, and the bound for the a_0 metric like $\lambda_1(A_\beta)^{5/3}$. We test whether the same holds for the stirrer potential in three dimensions. On the box $(-8, 8)^3$ of Section 5.2 the range of β is limited, since the Thomas–Fermi radius grows like $\beta^{1/5}$ and leaves $(-8, 8)^3$ near $\beta = 2 \times 10^4$. Thus we run the test on the larger box $(-16, 16)^3$ with the same 499^3 grid and the same initial value, and we vary only β , with $\beta = 1600, 6400$ and 102400 , a factor of 64. With $a \approx 10^{-3}$ as in Section 1.1, these values correspond to condensates of about 6.4×10^4 to 4.1×10^6 particles, inside the experimental range but well above the $\beta = 200$ used in Bao and Du (2004). Figures 3 and 4 show the error curves, one row per β .

Over that range $\lambda_1(A_\beta)$ grows by a factor of 4.68, and the iteration counts follow the predicted growth of the condition numbers closely. The count of the a_u flow grows by a factor of 2.74, against the factor $\lambda_1^{2/3} = 2.80$ of Corollary 3.11, and that of its conjugate gradient variant grows by 1.69, against the factor $\sqrt{2.80} = 1.67$ of a $\sqrt{\kappa}$ iteration count. The count of the modified a_0 flow grows faster than that of the a_u flow and reaches the iteration cap at the largest β , in line with the growth $\lambda_1^{5/3} = 13.1$ of the bound of Proposition 3.6. The count of the modified a_u flow grows like that of the a_u flow, with a smaller constant. The iteration count of the Thomas–Fermi metric does not grow. Its error curves have the same slope in the three rows, while every other curve flattens as β grows, in agreement with (29) and Remark 3.12. At the largest β the boundary layer is about $\lambda_1^{-1/6} = 0.46$ wide and is resolved by about seven cells of size $h = 0.064$. The resolution does not affect the comparison, since each run is measured against a reference on its own grid. The sweep thus measures the conditioning of the discrete problem in each metric, which is the quantity bounded in Section 3.

The modified H^1 flow with the fixed shift $\alpha = 20$ behaves unlike the other two Euclidean metrics. Its iteration count changes little with β and eventually falls below that of the a_0 flow. The three-dimensional sweep of Figure A.1 explains this through the negative part of Z . For the a_0 metric, ζ_- equals $\lambda_1(A_\beta)$ exactly in every run because the maximum in (20) is attained where the density vanishes, whereas for the modified H^1 metric ζ_- is at most α to leading order in the Thomas–Fermi approximation $\beta \mathbf{u}^2 \approx (\lambda_1(A_\beta) - \mathbb{V})_+$, because then $\mathbb{V} + 3\beta \mathbf{u}^2 \geq \lambda_1(A_\beta)$ everywhere. The measured values are 19.4, 18.2, 16.8 and 15.2 against $\alpha = 20$ at $\beta = 10^2, 10^3, 10^4$ and 10^5 . Since σ_T grows like $\lambda_1(A_\beta)^{1/3}$, the lower bound $\sigma_T / (\sigma_T + \zeta_-)$ of Theorem 3.4 tends to one for the modified H^1 metric and to zero for the a_0 metric, and the exact $\min \rho_{\mathbf{u}}$ in Figure A.1 rises for the modified H^1 metric and falls for the a_0 metric. The upper edge, however, is governed by ζ_+ , which is of the order of $\max_i \mathbb{V}_i$. This suggests that the condition number of the modified H^1 metric is determined by the size of the box rather than by β (Section 5.4). The condition numbers of the modified H^1 and a_0 metrics cross between $\beta = 10^3$ and 10^4 .

5.4. Dependence on the domain size

Remark 3.7 bounds the condition number of the modified a_0 metric uniformly in the size of the box for a fixed mesh size, provided the coercivity constant stays bounded below, while the bound for the modified H^1 metric grows with the box. We test domain robustness numerically with a fixed number of unknowns instead of a fixed mesh size, on nested boxes $L = 8, 16$ and 32 , with the stirrer potential (40),

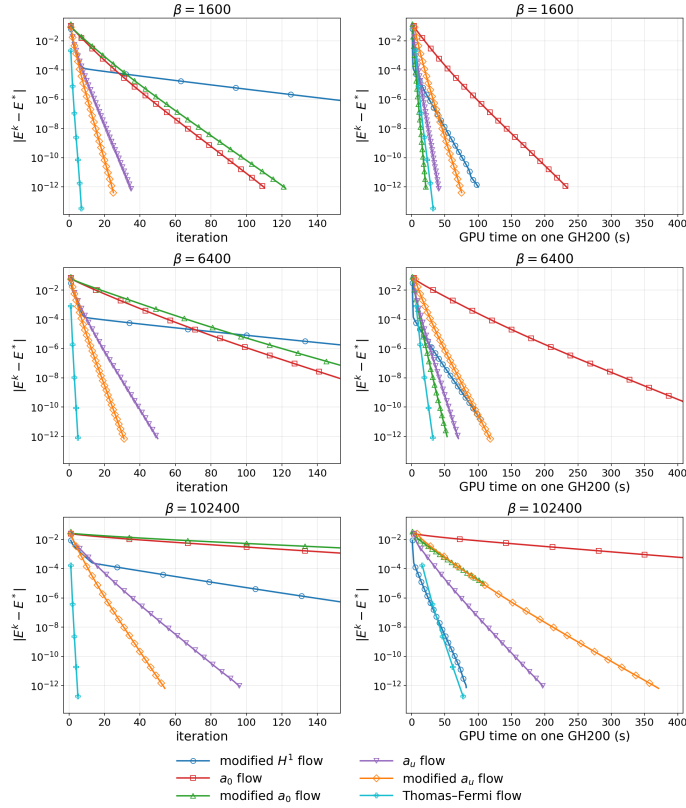


FIG. 3. The six gradient flows against the interaction strength, on $(-16, 16)^3$ at 499^3 , in the drawing convention of Figure 1, with $\beta = 1600, 6400$ and 102400 from top to bottom and axes cut at 150 iterations and 400 seconds. A curve that leaves a panel does not converge inside it.

whose harmonic part is confining as the remark requires, and with the same $\beta = 1600$, initial value, $\alpha = 20$, inner solver and 499^3 unknowns. The mesh is thus coarser on the larger boxes, with $h = 0.032, 0.064$ and 0.128 . The run with $L = 16$ is the $\beta = 1600$ run of Section 5.3. Figures 5 and 6 show the error curves.

For the five metrics built from the potential, the iteration counts are independent of L to within one iteration for four of them and to within two iterations for the Thomas–Fermi flow. The error curves of these five metrics in Figures 5 and 6 are also nearly independent of L . The iteration count of the modified H^1 flow grows with L , and at $L = 32$ it does not reach the tolerance within the iteration cap. Its conjugate gradient variant needs about twice as many iterations per doubling of L . Since one iteration costs the same in all three boxes, every method except the two using the modified H^1 metric takes about the same wall clock time at $L = 32$ as at $L = 8$, or less. With conjugate gradients, the modified a_0 metric is as fast as the modified H^1 metric on the smallest box and the fastest of the six conjugate gradient methods on the two larger boxes. The a_0 metric is as robust in iterations as its modified version and an order of magnitude more expensive at every L .

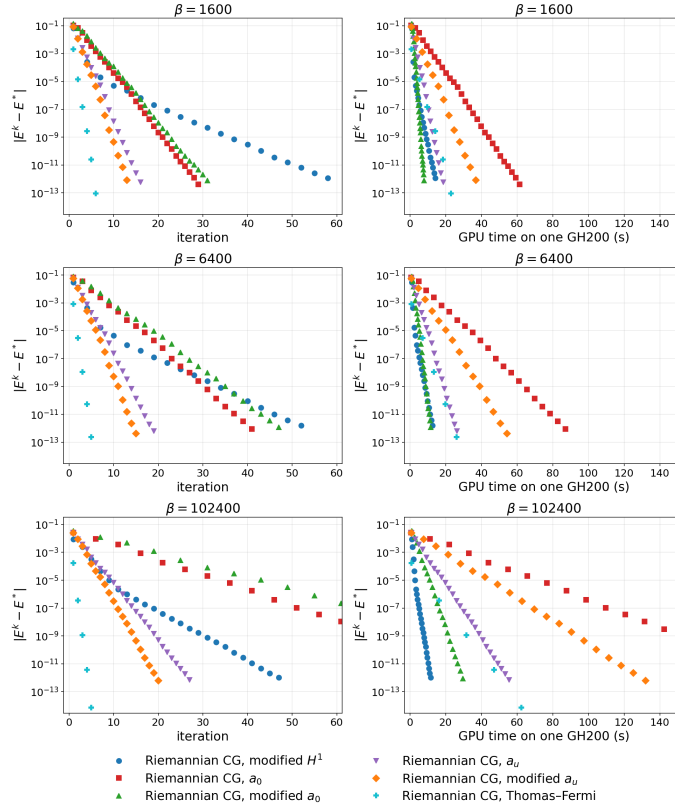


FIG. 4. The same for the six conjugate gradient methods, cut at 60 iterations and 150 seconds.

6. Conclusions

We have analyzed fully discretized Sobolev gradient flows for the Gross–Pitaevskii ground state problem through the condition number of the Riemannian Hessian at the ground state. The quadratic form of this Hessian does not depend on the metric, so bounds of its Rayleigh quotient give a uniform analysis of the flows for different metrics. The bound for the shifted Laplacian metric grows with the domain size, and so do its iteration counts in the tests. The condition number of the energy-adaptive metric grows like the two-thirds power of the ground state eigenvalue in the Thomas–Fermi regime, as proved for the one-dimensional harmonic trap and observed in two and three dimensions. Inspired by these bounds, we also consider three new metrics: a metric with only the separable part of the potential, which is robust in the domain size and inverted exactly by tensor-product solvers, a modified energy-adaptive metric that asymptotically halves the condition number, and a Thomas–Fermi metric whose condition number is bounded uniformly in the interaction strength, as proved for the one-dimensional harmonic trap and observed in three dimensions. For every metric, the local rate of the fixed-step iteration is determined by the spectrum of the Hessian, and the iterates converge to a single critical point from any initial value. The three-dimensional experiments are consistent with the analysis for both Riemannian gradient descent and Riemannian conjugate gradient methods. The Thomas–Fermi

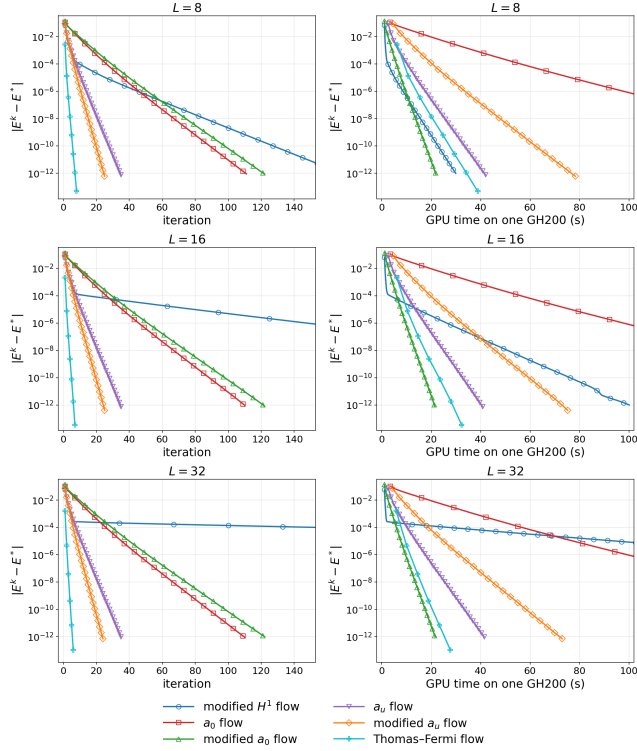


FIG. 5. The six gradient flows on the nested boxes $(-L, L)^3$, each on a 499^3 grid, in the drawing convention of Figure 1: the energy error against the iteration count on the left and against GPU time on the right, with $L = 8, 16$ and 32 from top to bottom. All three rows use the same axes, cut at 150 iterations and 100 seconds, and a curve that leaves a panel does not converge inside it.

flow needs the fewest iterations. The shifted Laplacian metric and the metric with only the separable part of the potential are the fastest in computational time, because their iterations are the cheapest.

Acknowledgements

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.

Funding

Xiangxiong Zhang was supported in part by NSF DMS-2208518.

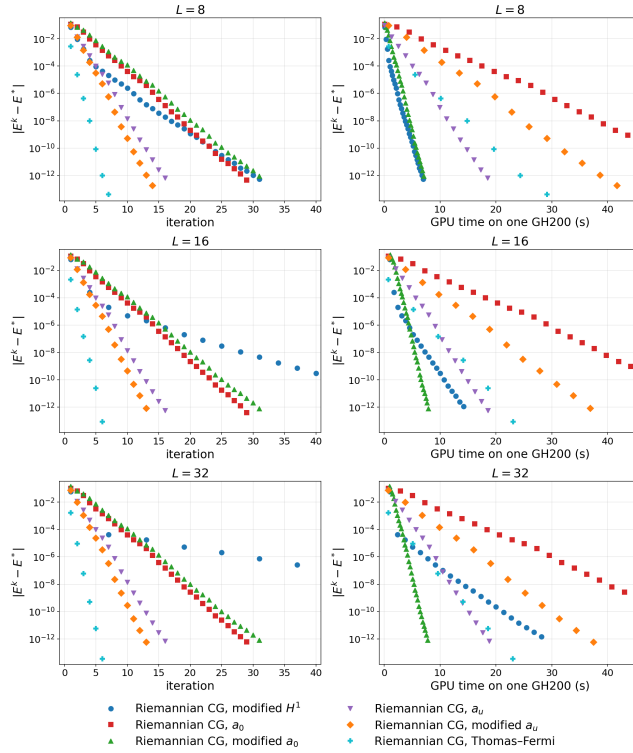


FIG. 6. The same for the six conjugate gradient methods, on the same boxes and with the same convention, cut at 40 iterations and 45 seconds.

Use 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.

Data Availability Statements

The codes of all numerical tests are available on <https://github.com/zhan1966/GPE/>.

REFERENCES

- P.-A. Absil, R. Mahony, and R. Sepulchre. *Optimization algorithms on matrix manifolds*. Princeton University Press, Princeton, NJ, 2008.
- R. Altmann, P. Henning, and D. Peterseim. The J-method for the Gross–Pitaevskii eigenvalue problem. *Numerische Mathematik*, 148(3):575–610, 2021.
- R. Altmann, D. Peterseim, and T. Stykel. Energy-adaptive Riemannian optimization on the Stiefel manifold. *ESAIM: Mathematical Modelling and Numerical Analysis*, 56(5):1629–1653, 2022.

- R. Altmann, D. Peterseim, and T. Stykel. Riemannian Newton methods for energy minimization problems of Kohn–Sham type. *Journal of Scientific Computing*, 101(1):6, 2024.
- R. Altmann, M. Hermann, D. Peterseim, and T. Stykel. Riemannian optimization methods for ground states of multicomponent Bose–Einstein condensates. *IMA Journal of Numerical Analysis*, 46(4):1994–2033, 2026.
- X. Antoine and R. Duboscq. Robust and efficient preconditioned Krylov spectral solvers for computing the ground states of fast rotating and strongly interacting Bose–Einstein condensates. *Journal of Computational Physics*, 258:509–523, 2014.
- X. Antoine, A. Levitt, and Q. Tang. Efficient spectral computation of the stationary states of rotating Bose–Einstein condensates by preconditioned nonlinear conjugate gradient methods. *Journal of Computational Physics*, 343: 92–109, 2017.
- W. Bao and Y. Cai. Mathematical theory and numerical methods for Bose–Einstein condensation. *Kinetic and Related Models*, 6(1):1–135, 2013.
- W. Bao and Q. Du. Computing the ground state solution of Bose–Einstein condensates by a normalized gradient flow. *SIAM Journal on Scientific Computing*, 25(5):1674–1697, 2004.
- N. Boumal. *An introduction to optimization on smooth manifolds*. Cambridge University Press, Cambridge, 2023.
- E. Cancès, R. Chakir, and Y. Maday. Numerical analysis of nonlinear eigenvalue problems. *Journal of Scientific Computing*, 45(1–3):90–117, 2010.
- H. Chen, G. Dong, W. Liu, and Z. Xie. Second-order flows for computing the ground states of rotating Bose–Einstein condensates. *Journal of Computational Physics*, 475:111872, 2023.
- Z. Chen, J. Lu, Y. Lu, and X. Zhang. On the convergence of Sobolev gradient flow for the Gross–Pitaevskii eigenvalue problem. *SIAM Journal on Numerical Analysis*, 62(2):667–691, 2024.
- Z. Chen, J. Lu, Y. Lu, and X. Zhang. Fully discretized Sobolev gradient flow for the Gross–Pitaevskii eigenvalue problem. *Mathematics of Computation*, 94(356):2723–2760, 2025.
- I. Danaila and P. Kazemi. A new Sobolev gradient method for direct minimization of the Gross–Pitaevskii energy with rotation. *SIAM Journal on Scientific Computing*, 32(5):2447–2467, 2010.
- I. Danaila and B. Protas. Computation of ground states of the Gross–Pitaevskii functional via Riemannian optimization. *SIAM Journal on Scientific Computing*, 39(6):B1102–B1129, 2017.
- E. Faou and T. Jézéquel. Convergence of a normalized gradient algorithm for computing ground states. *IMA Journal of Numerical Analysis*, 38(1):360–376, 2018.
- Z. Feng and Q. 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.
- Z. Feng, Q. Tang, and C. 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.
- C. Gallo and D. Pelinovsky. On the Thomas–Fermi ground state in a harmonic potential. *Asymptotic Analysis*, 73 (1–2):53–96, 2011.
- E. P. Gross. Structure of a quantized vortex in boson systems. *Il Nuovo Cimento*, 20(3):454–477, 1961.
- M. Hauck, Y. Liang, and D. Peterseim. Positivity preserving finite element method for the Gross–Pitaevskii ground state: discrete uniqueness and global convergence. *Numerische Mathematik*, 158(3):985–1014, 2026.
- P. 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(7):1517–1544, 2023.
- P. Henning and E. Jarlebring. The Gross–Pitaevskii equation and eigenvector nonlinearities: numerical methods and algorithms. *SIAM Review*, 67(2):256–317, 2025.
- P. Henning and A. Persson. On optimal convergence rates for discrete minimizers of the Gross–Pitaevskii energy in localized orthogonal decomposition spaces. *Multiscale Modeling & Simulation*, 21(3):993–1011, 2023.
- P. Henning and D. Peterseim. Sobolev gradient flow for the Gross–Pitaevskii eigenvalue problem: Global convergence and computational efficiency. *SIAM Journal on Numerical Analysis*, 58(3):1744–1772, 2020.
- P. Henning and M. Yadav. Convergence of a Riemannian gradient method for the Gross–Pitaevskii energy functional in a rotating frame. *ESAIM: Mathematical Modelling and Numerical Analysis*, 59(2):1145–1175,

- 2025a.
- P. Henning and M. Yadav. On discrete ground states of rotating Bose–Einstein condensates. *Mathematics of Computation*, 94(351):1–32, 2025b.
- P. Henning, A. Målqvist, and D. Peterseim. Two-level discretization techniques for ground state computations of Bose–Einstein condensates. *SIAM Journal on Numerical Analysis*, 52(4):1525–1550, 2014.
- G. Karali and C. Sourdis. The ground state of a Gross–Pitaevskii energy with general potential in the Thomas–Fermi limit. *Archive for Rational Mechanics and Analysis*, 217(2):439–523, 2015.
- P. Kazemi and M. Eckart. Minimizing the Gross–Pitaevskii energy functional with the Sobolev gradient—analytical and numerical results. *International Journal of Computational Methods*, 7(3):453–475, 2010.
- J. Li, S. Zheng, and X. Zhang. Global convergence of an efficient splitting method for the defocusing Gross–Pitaevskii ground state problem. *arXiv preprint arXiv:2609.13509*, 2026.
- E. H. Lieb, R. Seiringer, and J. Yngvason. Bosons in a trap: A rigorous derivation of the Gross–Pitaevskii energy functional. *Physical Review A*, 61(4):043602, 2000.
- X. Liu and X. Zhang. A simple GPU-accelerated solver for the Schrödinger operator with applications to ground states and Hamiltonian simulation. *arXiv preprint arXiv:2605.20491*, 2026.
- X. Liu, J. Shen, and X. 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(5):1157–1185, 2024.
- S. Łojasiewicz. Une propriété topologique des sous-ensembles analytiques réels. In *Les Équations aux Dérivées Partielles*, pages 87–89. Éditions du CNRS, Paris, 1963.
- J. W. Neuberger. *Sobolev Gradients and Differential Equations*, volume 1670 of *Lecture Notes in Mathematics*. Springer, Berlin, 1997.
- L. P. Pitaevskii. Vortex lines in an imperfect Bose gas. *Soviet Physics JETP*, 13(2):451–454, 1961.
- R. Schneider and A. Uschmajew. Convergence results for projected line-search methods on varieties of low-rank matrices via Łojasiewicz inequality. *SIAM Journal on Optimization*, 25(1):622–646, 2015.
- C. Zhang, P. Henning, M. Yadav, and W. Chen. Convergence analysis of Sobolev gradient flows for the rotating Gross–Pitaevskii energy functional. *arXiv preprint arXiv:2510.15604*, 2025.
- Z. Zhang. Exponential convergence of Sobolev gradient descent for a class of nonlinear eigenproblems. *Communications in Mathematical Sciences*, 20(2):377–403, 2022.
- S. Zheng, W. Huang, B. Vandereycken, and X. Zhang. Riemannian optimization using three different metrics for Hermitian PSD fixed-rank constraints. *Computational Optimization and Applications*, 91(3):1135–1184, 2025.

A. Exact spectra of the Riemannian Hessian

We test the spectral predictions of Section 3 on $(-L, L)^d$ with $V = |\mathbf{x}|^2$, using second-order finite differences and $\ell = 1/2$ in the Thomas–Fermi metric. In one dimension, $N = 3000$ for $\beta \leq 100$ and $N = 4000$ otherwise, with $L = \max\{6, 1.25(3\beta/4)^{1/3} + 4\}$. In two dimensions the grid is 200^2 with $L = \max\{6, (2\beta/\pi)^{1/4} + 3.5\}$, and in three dimensions it is 199^3 with $L = 1.25(15\beta/(8\pi))^{1/5} + 4$. All three values of L are derived from the Thomas–Fermi radius. At the largest β in three dimensions, the layer of width $\lambda_1^{-1/6}$ has about 2.5 grid points. Repeating the three-dimensional cases $\beta = 10^5$ on 399^3 and $\beta = 3 \times 10^5$ on 299^3 reproduces all condition numbers within 0.4%, except for the modified a_u metric.

Ground states are computed by the a_u flow with exact line search followed by Newton iterations until the residual reaches rounding error. Extreme Hessian eigenvalues are computed from the pencil $(\mathbb{M}H^*, \mathbb{M}B^*)$ on the tangent space, by a dense solver in one dimension, shift-invert Lanczos in two dimensions and locally optimal block preconditioned conjugate gradients with the preconditioner A_1^{-1}

β	$\lambda_1(A_\beta)$	$\lambda_2 - \lambda_1$	$\lambda_1(A_{3\beta}) - \lambda_1$	$\min \rho_{\mathbf{u}^*}$	$1 - \lambda_1/\lambda_2$	$1 - \lambda_1/\Lambda$	$\max \rho_{\mathbf{u}^*}$	$\kappa_{a_u}(\mathbf{u}^*)$ (bound)
$d = 1, N = 3000$ for $\beta \leq 10^2$ and $N = 4000$ otherwise								
1	1.383	1.82	0.734	0.6751	0.568	0.5680	1.0000	1.5 (2.5)
10^2	17.84	0.259	6.18	0.2806	1.4×10^{-2}	0.2573	1.7849	6.4 (7.7)
10^3	82.56	0.056	10.4	0.1167	6.8×10^{-4}	0.1123	1.9498	16.7 (17.8)
10^4	383.2	0.0121	17.5	0.04429	3.1×10^{-5}	0.04363	1.9890	44.9 (45.8)
3×10^5	3699	0.00125	37.2	0.009998	3.4×10^{-7}	0.009964	1.9989	199.9 (200.7)
$d = 2, 200^2$ grid								
10	3.241	1.49	1.85	0.5123	0.315	0.3637	1.0316	2.0 (4.6)
10^2	8.286	0.661	4.14	0.3839	7.4×10^{-2}	0.3336	1.4677	3.8 (5.8)
10^3	25.36	0.219	6.68	0.2251	8.6×10^{-3}	0.2084	1.8172	8.1 (9.6)
10^4	79.84	0.0695	10.1	0.1170	8.7×10^{-4}	0.1124	1.9456	16.6 (17.8)
$d = 3, 199^3$ grid								
10^2	5.932	1.10	2.77	0.4001	0.157	0.3187	1.2128	3.0 (5.8)
10^3	13.30	0.493	4.76	0.2972	3.6×10^{-2}	0.2635	1.6090	5.4 (7.5)
10^4	32.58	0.200	7.07	0.1911	6.1×10^{-3}	0.1782	1.8431	9.6 (11.2)
10^5	81.44	0.0797	10.0	0.1137	9.8×10^{-4}	0.1095	1.9403	17.1 (18.3)

TABLE A.1 *Exact extreme eigenvalues $\min \rho_{\mathbf{u}^*}$ and $\max \rho_{\mathbf{u}^*}$ of the Riemannian Hessian at the ground state under the a_u metric for the harmonic trap $V = |\mathbf{x}|^2$ in $d = 1, 2, 3$ dimensions, the lower bound $1 - \lambda_1(A_\beta)/\lambda_2(A_\beta)$ based on the spectral gap alone, the lower bound $1 - \lambda_1(A_\beta)/\Lambda$ of Corollary 3.5, and the exact condition number $\kappa_{a_u}(\mathbf{u}^*)$ with the bound of Corollary 3.5 in parentheses.*

in three dimensions. The code reproduces $\lambda_1/\lambda_2 = 0.99909$ and $\max_i |1 - \rho_i| = 0.94192$ from Henning (2023, Section 7.2) to five digits.

Table A.1 tests the bounds of Corollary 3.5 for the a_u metric in the harmonic trap $V = |\mathbf{x}|^2$ in one, two and three dimensions. For $\beta \geq 10^2$, the lower bound $1 - \lambda_1(A_\beta)/\Lambda$ is within 9% of the exact minimum in one dimension and within 21% in two and three dimensions, the maximum approaches the upper bound 2, and the bound on $\kappa_{a_u}(\mathbf{u}^*)$ is within 2% of the exact value for $\beta \geq 10^4$ in one dimension. For comparison, the table also lists the bound $1 - \lambda_1(A_\beta)/\lambda_2(A_\beta)$ that uses the spectral gap alone, which is known not to be sharp (Henning, 2023, Remark 3.3). For $\beta \geq 10^4$ this bound is smaller than the exact minimum by a factor between 31 and 3×10^4 , because in the Thomas–Fermi regime $\Lambda = \lambda_1(A_{3\beta})$, and $\lambda_1(A_{3\beta}) - \lambda_1(A_\beta)$ grows with β while the gap $\lambda_2(A_\beta) - \lambda_1(A_\beta)$ closes.

The table also confirms the asymptotics of Corollary 3.11. In one dimension, the slopes against λ_1 over $\beta \in [10^4, 3 \times 10^5]$ are -0.66 for $\min \rho_{\mathbf{u}^*}$, 0.333 for $\lambda_1(A_{3\beta}) - \lambda_1$ and -1.00 for the gap $\lambda_2(A_\beta) - \lambda_1$. The ratio $(\lambda_1(A_{3\beta}) - \lambda_1)/\lambda_1^{1/3}$ is 2.40, 2.41, 2.41 at $\beta = 10^3, 10^4, 3 \times 10^5$, against $\mu_1 = 2.4105$ from collocation of the Painlevé-II equation and diagonalization of M_0 . In two dimensions the ratio is 2.35 at $\beta = 10^4$.

Table A.2 compares the modified a_u and Thomas–Fermi metrics with the a_u metric in one dimension, for $V(x) = x^2$ and for $V(x) = x^2 + 60e^{-(x-1)^2}$, the one-dimensional analogue of the stirrer potential (40) of Section 5. At $\beta = 10^2, 10^3, 10^4$ in the harmonic trap, gradient descent with exact line search on a grid with $N = 2000$, started from the Thomas–Fermi profile and stopped at $E(\mathbf{u}_k) - E(\mathbf{u}^*) < 10^{-12}$, takes 21, 47, 114 iterations under the modified a_u metric against 32, 84, 217

$V(x)$	β	λ_1	$e_{\text{TF}}/\lambda_1^{1/3}$	$\sigma_T/\lambda_1^{1/3}$	a_u	$\tilde{a}_u$	TF
x^2	10^2	17.84	0.39	2.37	6.4	3.9	1.17
x^2	10^3	82.56	0.41	2.40	16.7	8.9	1.16
x^2	10^4	383.16	0.42	2.41	44.9	22.5	1.16
$x^2 + 60e^{-(x-1)^2}$	10^2	24.93	0.95	2.38	6.7	4.6	1.18
$x^2 + 60e^{-(x-1)^2}$	1600	117.88	0.41	2.40	20.4	11.0	1.16
$x^2 + 60e^{-(x-1)^2}$	10^4	385.87	0.42	2.41	44.7	22.6	1.16

TABLE A.2 *Exact condition numbers $\kappa(\mathbf{u}^*)$ of the Riemannian Hessian at the ground state in one dimension (last three columns) under the a_u , modified a_u and Thomas–Fermi ($\ell = 1/2$) metrics, with the Thomas–Fermi error (27) and the coercivity constant σ_T in units of the layer scale $\lambda_1^{1/3}$, for the harmonic trap and for the trap with the stirrer bump.*

under the a_u metric, but it costs 42, 94, 228 solves against 32, 84, 217. The halved condition number thus reduces iterations without reducing solves.

For the Thomas–Fermi metric, the harmonic-trap data verify the two hypotheses of (29), with $e_{\text{TF}}/\lambda_1^{1/3} \rightarrow 0.43$ from below and $\sigma_T/\lambda_1^{1/3} \rightarrow \mu_1$. At $\beta = 10^4$, (28) gives $\kappa_{\text{TF}}(\mathbf{u}^*) \leq 1.38$ against the measured 1.16, a factor 39 smaller than $\kappa_{a_u}(\mathbf{u}^*)$. With the nonseparable bump, $\kappa_{\text{TF}}(\mathbf{u}^*)$ stays between 1.16 and 1.18. Figure A.1 extends the comparison to three dimensions over $\beta \in [1, 3 \times 10^5]$. There $\kappa_{\text{TF}}(\mathbf{u}^*)$ falls from 1.89 to 1.17, while $\kappa_{a_0}(\mathbf{u}^*)$ grows by a factor 330, and the measured exponents against λ_1 are 0.62, 1.62 and -0.12 for the a_u , a_0 and Thomas–Fermi metrics, consistent with the predicted $2/3$, with the $5/3$ in the a_0 bound and with (29).

The upper edge $s_{N-1}/(s_{N-1} + \lambda_1)$ of the modified a_u spectrum is mesh dependent by construction, and refinement changes $\kappa_{\tilde{a}_u}(\mathbf{u}^*)$ at $\beta = 10^4$ from 21.5 with $N = 2000$ to 22.5 with $N = 4000$ in one dimension and by at most 4% in the three-dimensional mesh checks. The lower edge is stable under refinement.

B. Riemannian connection, Hessian and vector transport

We work on an open set where $B(\mathbf{u})$ is smooth and $\mathbb{M}B(\mathbf{u})$ is symmetric positive definite. Write $\mathbb{G}(\mathbf{u}) := B(\mathbf{u})^{-1}$ and let $\odot$ denote componentwise multiplication. For $\mathbf{u} \in \mathcal{M}$, the $g_{\mathbf{u}}$ -orthogonal projection $P_{\mathbf{u}}$ is defined in (15), and by (9) and (14)

$$g_{\mathbf{u}}(\mathbb{G}(\mathbf{u})\mathbf{x}, \mathbf{y}) = \langle \mathbf{x}, \mathbf{y} \rangle_h, \quad N_{\mathbf{u}}\mathcal{M} = \text{span}\{\mathbb{G}(\mathbf{u})\mathbf{u}\}. \quad (\text{B.1})$$

The five metrics of (10) other than the Thomas–Fermi metric have the form

$$B(\mathbf{u}) = B_0 + c\beta \text{diag}(\mathbf{u}^2), \quad B_0 \text{ independent of } \mathbf{u}. \quad (\text{B.2})$$

Here $c = 0$ for the three Euclidean metrics, $c = 1$ for the a_u metric and $c = 3$ for the modified a_u metric, and $B_0 = -\Delta_h + \mathbb{V}$ gives the family (24). The coefficient c enters only the metric and not the energy (7).

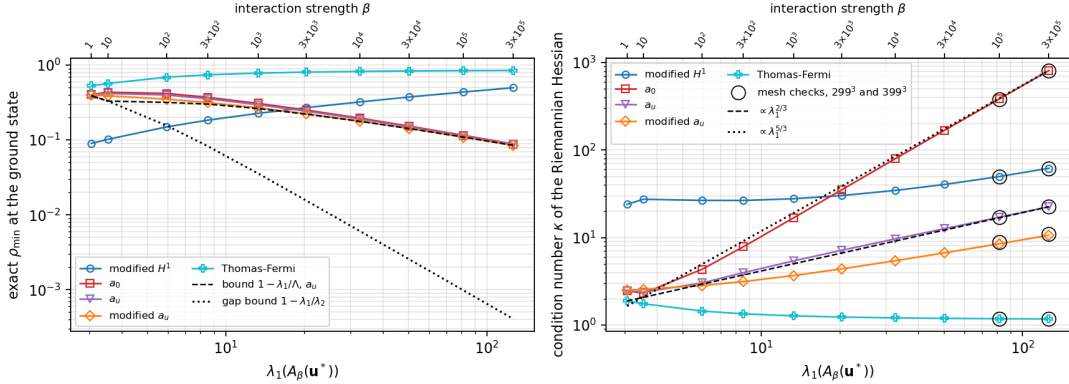


FIG. A.1. Three-dimensional harmonic trap on a 199^3 grid. The open rings are the two mesh checks, $\beta = 10^5$ on 399^3 and $\beta = 3 \times 10^5$ on 299^3 . Both panels are drawn against $\lambda_1(A_\beta(\mathbf{u}^*))$, with β on the top axis. Left: the exact $\min \rho_{\mathbf{u}^*}$ for the six metrics of Table 1, with $\ell = 1/2$ in the Thomas–Fermi metric, together with the lower bound $1 - \lambda_1(A_\beta)/\Lambda$ of Corollary 3.5 and the bound based on the spectral gap alone, both for the a_u metric. The modified a_0 metric coincides with the a_0 metric because $\mathbb{V}$ is separable here. Right: the exact condition numbers against the reference slopes $\lambda_1(A_\beta)^{2/3}$ and $\lambda_1(A_\beta)^{5/3}$.

LEMMA B.1 (Riemannian connection) For a smooth metric B , the ambient Levi–Civita connection is $\bar{\nabla}_\xi \chi = D\chi(\mathbf{u})[\xi] + \Gamma_{\mathbf{u}}(\xi, \chi)$, where the bilinear map $\Gamma_{\mathbf{u}}$ is determined by

$$\begin{aligned} 2 \langle B\Gamma_{\mathbf{u}}(\xi, \chi), \zeta \rangle_h &= \langle DB[\xi]\chi, \zeta \rangle_h + \langle DB[\chi]\xi, \zeta \rangle_h \\ &\quad - \langle DB[\zeta]\xi, \chi \rangle_h \end{aligned} \quad (\text{B.3})$$

for all $\zeta \in \mathbb{R}^N$, with all vector-field values and coefficients evaluated at $\mathbf{u}$. For (B.2),

$$\bar{\nabla}_\xi \chi = D\chi(\mathbf{u})[\xi] + c\beta \mathbb{G}(\mathbf{u})(\mathbf{u} \odot \xi \odot \chi). \quad (\text{B.4})$$

The induced connection on $\mathcal{M}$ is $\nabla_\xi \chi = P_{\mathbf{u}}(\bar{\nabla}_\xi \chi)$ for tangent vector fields.

Proof The Koszul formula (Absil et al., 2008, (5.11)) for constant vector fields gives (B.3), and the term $D\chi(\mathbf{u})[\xi]$ accounts for general vector fields (Absil et al., 2008, (5.7)). For (B.2), $DB[\xi] = 2c\beta \text{diag}(\mathbf{u} \odot \xi)$. Since $\mathbb{M}$ is diagonal, each inner product on the right of (B.3) equals $2c\beta \sum_i w_i u_i \xi_i \chi_i \zeta_i$. Their signed sum gives (B.4). Orthogonal projection gives the induced connection (Absil et al., 2008, Prop. 5.3.2). $\square$

LEMMA B.2 (Riemannian Hessian at an arbitrary point) Let $\mathbf{u} \in \mathcal{M}$, $\eta \in T_{\mathbf{u}}\mathcal{M}$, and $p_{\mathbf{u}} := \text{grad } E(\mathbf{u}) = \mathbb{G}(A_\beta(\mathbf{u})\mathbf{u} - \lambda_B(\mathbf{u})\mathbf{u})$, with λ_B as in (15). For a smooth metric,

$$\text{Hess } E(\mathbf{u})[\eta] = P_{\mathbf{u}} \left(\mathbb{G}(A_{3\beta}(\mathbf{u}) - \lambda_B(\mathbf{u})\mathbb{I})\eta - \mathbb{G}DB[\eta]p_{\mathbf{u}} + \Gamma_{\mathbf{u}}(\eta, p_{\mathbf{u}}) \right). \quad (\text{B.5})$$

For (B.2),

$$\text{Hess } E(\mathbf{u})[\eta] = P_{\mathbf{u}} \mathbb{G} \left((A_{3\beta}(\mathbf{u}) - \lambda_B(\mathbf{u})\mathbb{I})\eta - c\beta \mathbf{u} \odot \eta \odot p_{\mathbf{u}} \right). \quad (\text{B.6})$$

At a critical point, $p_{\mathbf{u}} = 0$ and $\lambda_B(\mathbf{u}) = \lambda(\mathbf{u})$, so for every smooth metric

$$g_{\mathbf{u}}(\text{Hess}E(\mathbf{u})[\eta], \eta) = \langle (A_{3\beta}(\mathbf{u}) - \lambda(\mathbf{u})\mathbb{I})\eta, \eta \rangle_h,$$

which recovers Lemma 3.1.

Proof Differentiating $p_{\mathbf{u}}$ along a curve on $\mathcal{M}$ with initial velocity η gives

$$Dp_{\mathbf{u}}[\eta] = \mathbb{G}(A_{3\beta}(\mathbf{u}) - \lambda_B(\mathbf{u})\mathbb{I})\eta - \mathbb{G}DB[\eta]p_{\mathbf{u}} - D\lambda_B(\mathbf{u})[\eta]\mathbb{G}\mathbf{u}.$$

The last term is normal and vanishes under $P_{\mathbf{u}}$. By Absil et al. (2008, Def. 5.5.1 and Prop. 5.3.2), $\text{Hess}E(\mathbf{u})[\eta]$ is the projection $P_{\mathbf{u}}$ of the ambient covariant derivative of the vector field p along η , so adding the connection term and projecting proves (B.5), and (B.4) gives (B.6). At a critical point the metric-derivative terms vanish, and (B.1) gives the quadratic-form identity. Away from critical points, $\lambda_B(\mathbf{u})$ cannot in general be replaced by $\lambda(\mathbf{u})$. $\square$

The Thomas–Fermi metric. The operator $B_{\text{TF}}(\mathbf{u})$ of (11) depends on $\mathbf{u}$ also through $\lambda(\mathbf{u})$, so it is not of the form (B.2). It is smooth on the open set of all $\mathbf{u} \neq 0$ with $\lambda(\mathbf{u}) \neq \nabla_i$ for every node i . On this set, Lemmas B.1 and B.2 apply with the general formulas (B.3) and (B.5).

Retraction and vector transport. The projection transport used in Section 5.1 is

$$\mathcal{T}_{\eta}(\xi) := P_{\mathbf{v}}\xi = \xi - \frac{\langle \mathbf{v}, \xi \rangle_h}{\langle \mathbf{v}, \mathbb{G}(\mathbf{v})\mathbf{v} \rangle_h} \mathbb{G}(\mathbf{v})\mathbf{v}, \quad \mathbf{v} = R_{\mathbf{u}}(\eta), \quad (\text{B.7})$$

for $\eta, \xi \in T_{\mathbf{u}}\mathcal{M}$. It is the transport (Absil et al., 2008, (8.10)) with the g -orthogonal projection in place of the Euclidean one. It lies in $T_{\mathbf{v}}\mathcal{M}$, is linear in ξ , satisfies $\mathcal{T}_0(\xi) = \xi$ and depends smoothly on $(\mathbf{u}, \eta, \xi)$ wherever B is smooth. These are the defining properties of a vector transport associated with R (Absil et al., 2008, Def. 8.1.1). Differentiating the retraction gives another transport (Absil et al., 2008, (8.6) and Example 8.1.4),

$$DR_{\mathbf{u}}(\eta)[\xi] = \frac{\xi - \mathbf{v}\langle \mathbf{v}, \xi \rangle_h}{\sqrt{1 + \|\eta\|_h^2}}, \quad (\text{B.8})$$

which differs from (B.7) in general, and neither one is in general the parallel transport of the Riemannian connection. In the metric norm $\|\xi\|_{\mathbf{u}} := g_{\mathbf{u}}(\xi, \xi)^{1/2}$ of Section 4, orthogonal projection gives $\|\mathcal{T}_{\eta}(\xi)\|_{\mathbf{v}} \leq \|\xi\|_{\mathbf{v}}$, but in general not $\|\mathcal{T}_{\eta}(\xi)\|_{\mathbf{v}} \leq \|\xi\|_{\mathbf{u}}$ when the metric depends on $\mathbf{u}$. Lemma 3.1 holds along every curve, in particular along $t \mapsto R_{\mathbf{u}}(t\eta)$, because the acceleration term is multiplied by $\text{grad}E(\mathbf{u}) = 0$ at a critical point, so no second-order property of R is needed.