ESAIM: M2AN 56 (2022) 1629–1653
https://doi.org/10.1051/m2an/2022036

ESAIM: Mathematical Modelling and Numerical Analysis
www.esaim-m2an.org

ENERGY-ADAPTIVE RIEMANNIAN OPTIMIZATION ON THE STIEFEL MANIFOLD

Robert Altmann[1,*], Daniel Peterseim[2] and Tatjana Stykel[2]

Abstract. This paper addresses the numerical solution of nonlinear eigenvector problems such as the Gross–Pitaevskii and Kohn–Sham equation arising in computational physics and chemistry. These problems characterize critical points of energy minimization problems on the infinite-dimensional Stiefel manifold. To efficiently compute minimizers, we propose a novel Riemannian gradient descent method induced by an energy-adaptive metric. Quantified convergence of the methods is established under suitable assumptions on the underlying problem. A non-monotone line search and the inexact evaluation of Riemannian gradients substantially improve the overall efficiency of the method. Numerical experiments illustrate the performance of the method and demonstrates its competitiveness with well-established schemes.

Mathematics Subject Classification. 65N25, 81Q10.

Received August 23, 2021. Accepted April 8, 2022.

1. Introduction

This paper is devoted to the numerical solution of energy minimization problems stated on the infinite-dimensional Stiefel manifold of index N containing N-tuples of L^2-orthonormal functions. The Kohn–Sham model [19, 23, 25] is a prototypical example. In this popular model from *density functional theory* in computational chemistry, the state of the system is described by $N > 1$ functions (orbitals), which need to satisfy L^2-orthogonality conditions. The ground state of the system minimizes the Kohn–Sham energy under these orthogonality constraints, *i.e.*, on the Stiefel manifold of index N, *cf.* [36]. For $N = 1$, the Stiefel manifold boils down to the unit sphere in L^2. In this special case, the Gross–Pitaevskii model for Bose–Einstein condensates of ultracold bosonic gases [26, 29] is a relevant example. Its ground state is the global minimizer of the corresponding Gross–Pitaevskii energy functional on the Stiefel manifold which simply represents a unit mass constraint.

More generally, the ground states of energy functionals on the Stiefel manifold as well as further critical points are characterized by coupled systems of eigenvalue problems of partial differential equations (PDEs) with eigenvector nonlinearities, so-called *nonlinear eigenvector problems*. Existing approximation methods for

Keywords and phrases. Riemannian optimization, Stiefel manifold, Kohn–Sham model, Gross–Pitaevskii eigenvalue problem, nonlinear eigenvector problem.

[1] Institute of Mathematics, University of Augsburg, Universitätsstr. 12a, 86159 Augsburg, Germany.

[2] Institute of Mathematics & Centre for Advanced Analytics and Predictive Sciences (CAAPS), University of Augsburg, Universitätsstr. 12a, 86159 Augsburg, Germany.

*Corresponding author: robert.altmann@math.uni-augsburg.de

these problems are either linked to linear eigenvalue solvers or to Riemannian optimization. A well-known iteration scheme for the nonlinear eigenvector problem is the *self-consistent field iteration* (SCF). Each SFC iteration step involves the solution of a *linear* eigenvalue problem, see, *e.g.*, [9,10,13] and [21] for its connection to Newton's method. On the Riemmanian side, the *direct constrained minimization algorithm* (DCM) is very popular. DCM results from a standard minimization approach [3,31,35] and is based on the Riemannian gradient descent method in L^2. However, this method requires suitable preconditioning to work. In the special case of the Gross–Pitaevskii eigenvalue problem, the DCM is known as the *discrete normalized gradient flow* [7]. Although empirically successful, the preconditioning or stable time discretization comes with the drawback of deviating from the gradient descent structure. In this case, the energy decay cannot be guaranteed anymore. In [18], an alternative Riemannian gradient descent scheme was proposed for the special case of the Gross–Pitaevskii problem, which is based on a gradient flow defined in an energy-adaptive metric. The resulting method is convergent and energy diminishing for sufficiently small step sizes. The energy diminishing property even gives rise to global convergence to the ground state [18] and turns out to be valuable in the context of reliable *a posteriori* error control [17].

In this paper, we generalize this promising yet simple energy-adaptive Riemannian descent method to nonlinear eigenvector problems formulated on the Stiefel manifold. The general functional analytical setting of the considered problems is presented in Section 2. Details on the infinite-dimensional Stiefel manifold, its tangent and normal spaces, and the orthogonal projection onto the tangent space are then discussed in Section 3. Therein, we show that the mentioned projection can be characterized by a saddle point problem, which facilitates the proposed algorithm significantly. Finally, several retractions are introduced, which are needed to transform tangent vectors back to the manifold. Section 4 presents the novel energy-adaptive Riemannian gradient descent method. Its convergence analysis generalizes the approach of [38] for $N = 1$. It is independent of the space dimension and, hence, also independent of possible spatial discretization by finite elements, spectral methods or related schemes. The convergence is further accelerated by the non-monotone line search algorithm of [34,39]. Moreover, we identify a connection to a preconditioned version of DCM, which motivates the substantial reduction of the computational complexity of the new method related on inexact gradient computations. In Section 5, we show that the Gross–Pitaevskii and Kohn–Sham models fit into the given framework. Numerical experiments for the Kohn–Sham model illustrate the performance of the presented method. Using the step size control and suitable inexact gradient computations prove the new approach competitive with established methods such as SCF and DCM.

2. Energy minimization problem on the Stiefel manifold

This section introduces an abstract constrained PDE energy minimization problem and its connection to a coupled system of nonlinear eigenvector problems formulated on the infinite-dimensional Stiefel manifold of index N. Particular examples such as the Gross–Pitaevskii eigenvalue problem and the Kohn–Sham model will be discussed in detail in Section 5.

2.1. Spaces and bilinear forms

We consider a space $\tilde{V} \subseteq H^1(\Omega)$ for a given domain $\Omega \subseteq \mathbb{R}^d$ and define

$$V := \tilde{V}^N, \qquad H := \big[L^2(\Omega)\big]^N$$

with $N \geq 1$. The suitable choice of the Hilbert space $\tilde{V}$ depends on the particular application, *cf.* the examples in Section 5. Let V^* denote the dual space of V. We assume that $V \subset H \subset V^*$ form a Gelfand triple ([37], Chap. 23.4). Throughout this paper, we use the row-vector notation for N-frames, *i.e.*, we write $\boldsymbol{v} = (v_1, \dots, v_N) \in V$. This allows us to adapt the notion of typical matrix-vector multiplication, *i.e.*, we may multiply $\boldsymbol{v}$ by an $N \times N$ matrix from the right, leading again to an element of V. Furthermore, for $\boldsymbol{v}, \boldsymbol{w} \in H$, we

define the dot product $\boldsymbol{v}\cdot\boldsymbol{w} := \sum_{j=1}^{N} v_j w_j$. We say that the components of $\boldsymbol{v} \in V\setminus\{\boldsymbol{0}\}$ are linearly independent, if there is no non-zero vector $x \in \mathbb{R}^N$ such that $\boldsymbol{v}x = 0$.

On the pivot space H, we introduce an outer product $[\![\cdot,\cdot]\!]_H \colon H\times H \to \mathbb{R}^{N\times N}$ and an inner product $(\cdot,\cdot)_H \colon H\times H\to\mathbb{R}$. More precisely, for $\boldsymbol{v},\boldsymbol{w}\in H$, we define

$$[\![\boldsymbol{v},\boldsymbol{w}]\!]_H := \begin{bmatrix} (v_1,w_1)_{L^2(\Omega)} & \dots & (v_1,w_N)_{L^2(\Omega)} \\ \vdots & \ddots & \vdots \\ (v_N,w_1)_{L^2(\Omega)} & \dots & (v_N,w_N)_{L^2(\Omega)} \end{bmatrix} \in \mathbb{R}^{N\times N} \tag{2.1}$$

and

$$(\boldsymbol{v},\boldsymbol{w})_H := \sum_{j=1}^{N} (v_j,w_j)_{L^2(\Omega)} = \operatorname{tr}[\![\boldsymbol{v},\boldsymbol{w}]\!]_H, \tag{2.2}$$

where tr denotes the trace of a matrix. The inner product (2.2) induces the norm $\|\boldsymbol{v}\|_H = \sqrt{(\boldsymbol{v},\boldsymbol{v})_H}$ on H. Some properties of the outer product (2.1) are collected in the following lemma, which follows from straight-forward calculations.

Lemma 2.1. *Consider $\boldsymbol{v},\boldsymbol{w}\in H$ and an arbitrary matrix $S\in\mathbb{R}^{N\times N}$. Then it holds that*

$$[\![\boldsymbol{v},\boldsymbol{w}S]\!]_H = [\![\boldsymbol{v},\boldsymbol{w}]\!]_H S, \qquad [\![\boldsymbol{v}S,\boldsymbol{w}]\!]_H = S^T[\![\boldsymbol{v},\boldsymbol{w}]\!]_H, \qquad [\![\boldsymbol{v},\boldsymbol{w}]\!]_H = [\![\boldsymbol{w},\boldsymbol{v}]\!]_H^T.$$

For the definition of the energy in the next subsection, we further introduce a (problem-dependent) bilinear form $a_{\boldsymbol{\phi}} \colon V\times V\to\mathbb{R}$ for a fixed $\boldsymbol{\phi}\in V$. With the density function $\rho(\boldsymbol{\phi}) = \boldsymbol{\phi}\cdot\boldsymbol{\phi}$, we consider

$$a_{\boldsymbol{\phi}}(\boldsymbol{v},\boldsymbol{w}) = a_0(\boldsymbol{v},\boldsymbol{w}) + \int_\Omega \gamma(\rho(\boldsymbol{\phi}))\,\boldsymbol{v}\cdot\boldsymbol{w}\,\mathrm{d}r = \sum_{j=1}^{N} \tilde{a}_{\boldsymbol{\phi}}(v_j,w_j) \tag{2.3}$$

for $\boldsymbol{v},\boldsymbol{w}\in V$. Here, $a_0 : V\times V\to\mathbb{R}$ is a bilinear form, which is independent of $\boldsymbol{\phi}$, and $\gamma:\mathbb{R}\to\mathbb{R}$ is a continuous nonlinear function with $\gamma(0)=0$. Later, a_0 and the term with γ will correspond, respectively, to the quadratic part and the nonlinear part of the energy. Note that (2.3) encodes a special structure, *i.e.*, $a_{\boldsymbol{\phi}}$ can be written as a sum with a bilinear form $\tilde{a}_{\boldsymbol{\phi}} \colon \tilde{V}\times\tilde{V}\to\mathbb{R}$. Within the abstract setting, we consider the following assumption.

Assumption 2.2 (Bilinear form $\tilde{a}_{\boldsymbol{\phi}}$)**.** *For a fixed $\boldsymbol{\phi}\in V$, $\tilde{a}_{\boldsymbol{\phi}}$ from (2.3) is a symmetric, bounded, and coercive bilinear form on $\tilde{V}$.*

By equation (2.3), the bilinear from $a_{\boldsymbol{\phi}}$ inherits the inner product structure from $\tilde{a}_{\boldsymbol{\phi}}$, meaning that $a_{\boldsymbol{\phi}}$ is symmetric, bounded, and coercive on V. Thus, it defines an inner product on V which induces the norm

$$\|\boldsymbol{v}\|_{a_{\boldsymbol{\phi}}} = \sqrt{a_{\boldsymbol{\phi}}(\boldsymbol{v},\boldsymbol{v})}, \qquad \boldsymbol{v}\in V.$$

The assumed Gelfand structure implies the existence of a constant $C_H > 0$ such that $\|\boldsymbol{v}\|_H \le C_H\|\boldsymbol{v}\|_{a_0}$. Moreover, for a bounded $\boldsymbol{\phi}\in V\cap[L^\infty(\Omega)]^N$, there exists a constant $c_E>0$ such that

$$c_E\|\boldsymbol{v}\|_{a_{\boldsymbol{\phi}}} \le \|\boldsymbol{v}\|_{a_0} \le \|\boldsymbol{v}\|_{a_{\boldsymbol{\phi}}} \qquad \text{for all } \boldsymbol{v}\in V.$$

The corresponding operator formulation of the bilinear form $a_{\boldsymbol{\phi}}$ reads

$$\langle \mathcal{A}_{\boldsymbol{\phi}}\boldsymbol{v},\boldsymbol{w}\rangle := a_{\boldsymbol{\phi}}(\boldsymbol{v},\boldsymbol{w}) \qquad \text{for all } \boldsymbol{v},\boldsymbol{w}\in V, \tag{2.4}$$

with a linear operator $\mathcal{A}_{\boldsymbol{\phi}} \colon V\to V^*$. Assumption 2.2 implies that $\mathcal{A}_{\boldsymbol{\phi}}$ is symmetric, bounded, and coercive. Hence, it is invertible (for fixed $\boldsymbol{\phi}$). Its inverse satisfies

$$a_{\boldsymbol{\phi}}\left(\mathcal{A}_{\boldsymbol{\phi}}^{-1}\boldsymbol{v},\boldsymbol{w}\right) = (\boldsymbol{v},\boldsymbol{w})_H \qquad \text{for all } \boldsymbol{v},\boldsymbol{w}\in V. \tag{2.5}$$

Next, we show some useful properties of the matrix $\left[\!\left[\boldsymbol{v},\mathcal{A}_{\boldsymbol{\phi}}^{-1}\boldsymbol{v}\right]\!\right]_H$.

Proposition 2.3. *Let $\boldsymbol{\phi}, \boldsymbol{v} \in V$ and let $\mathcal{A}_{\boldsymbol{\phi}}$ be defined as in (2.4). Then, under Assumption 2.2, the matrix $\left[\!\left[\boldsymbol{v}, \mathcal{A}_{\boldsymbol{\phi}}^{-1}\boldsymbol{v}\right]\!\right]_H \in \mathbb{R}^{N\times N}$ is symmetric positive semidefinite. If, additionally, $\boldsymbol{v} \neq \boldsymbol{0}$ and its components are linearly independent, then $\left[\!\left[\boldsymbol{v}, \mathcal{A}_{\boldsymbol{\phi}}^{-1}\boldsymbol{v}\right]\!\right]_H$ is positive definite.*

Proof. Due to the additive structure of (2.3), there exists a symmetric and coercive operator $\tilde{\mathcal{A}}_{\boldsymbol{\phi}}\colon \tilde{V} \to \tilde{V}^*$ corresponding to the bilinear form $\tilde{a}_{\boldsymbol{\phi}}$ such that

$$\langle \mathcal{A}_{\boldsymbol{\phi}}\boldsymbol{v}, \boldsymbol{w}\rangle = \sum_{j=1}^{N} \left\langle \tilde{\mathcal{A}}_{\boldsymbol{\phi}} v_j, w_j \right\rangle \qquad \text{for all } \boldsymbol{v}, \boldsymbol{w} \in V.$$

Thus, we conclude that $\mathcal{A}_{\boldsymbol{\phi}}^{-1}\boldsymbol{v} = \left(\tilde{\mathcal{A}}_{\boldsymbol{\phi}}^{-1}v_1, \dots, \tilde{\mathcal{A}}_{\boldsymbol{\phi}}^{-1}v_N\right) \in V$. Moreover, since $\tilde{\mathcal{A}}_{\boldsymbol{\phi}}$ is symmetric, so is its inverse, which implies

$$\begin{aligned}\left(\left[\!\left[\boldsymbol{v}, \mathcal{A}_{\boldsymbol{\phi}}^{-1}\boldsymbol{v}\right]\!\right]_H\right)_{ij} &= \left(v_i, (\mathcal{A}_{\boldsymbol{\phi}}^{-1}\boldsymbol{v})_j\right)_{L^2(\Omega)} = \left(v_i, \tilde{\mathcal{A}}_{\boldsymbol{\phi}}^{-1}v_j\right)_{L^2(\Omega)} \\ &= \left(v_j, \tilde{\mathcal{A}}_{\boldsymbol{\phi}}^{-1}v_i\right)_{L^2(\Omega)} = \left(v_j, (\mathcal{A}_{\boldsymbol{\phi}}^{-1}\boldsymbol{v})_i\right)_{L^2(\Omega)} = \left(\left[\!\left[\boldsymbol{v}, \mathcal{A}_{\boldsymbol{\phi}}^{-1}\boldsymbol{v}\right]\!\right]_H\right)_{ji}.\end{aligned}$$

Further, for an arbitrary vector $x \in \mathbb{R}^N$, we get

$$\begin{aligned}x^T\left(\left[\!\left[\boldsymbol{v}, \mathcal{A}_{\boldsymbol{\phi}}^{-1}\boldsymbol{v}\right]\!\right]_H x\right) &= \sum_{i=1}^{N}\sum_{j=1}^{N}\left(v_i, \tilde{\mathcal{A}}_{\boldsymbol{\phi}}^{-1}v_j\right)_{L^2(\Omega)} x_i\, x_j = \left(\boldsymbol{v}x, \tilde{\mathcal{A}}_{\boldsymbol{\phi}}^{-1}(\boldsymbol{v}x)\right)_{L^2(\Omega)} \\ &= \tilde{a}_{\boldsymbol{\phi}}\left(\tilde{\mathcal{A}}_{\boldsymbol{\phi}}^{-1}(\boldsymbol{v}x), \tilde{\mathcal{A}}_{\boldsymbol{\phi}}^{-1}(\boldsymbol{v}x)\right) \geq 0.\end{aligned}$$

This shows that $\left[\!\left[\boldsymbol{v}, \mathcal{A}_{\boldsymbol{\phi}}^{-1}\boldsymbol{v}\right]\!\right]_H$ is positive semidefinite. Finally, if $\boldsymbol{v} \neq \boldsymbol{0}$ has linearly independent components, then for all $x \in \mathbb{R}^N \setminus \{0\}$, we have $\boldsymbol{v}x \neq 0$ and, hence, $\left[\!\left[\boldsymbol{v}, \mathcal{A}_{\boldsymbol{\phi}}^{-1}\boldsymbol{v}\right]\!\right]_H$ is positive definite. □

2.2. Variational form and nonlinear eigenvector problem

Given an index $N \in \mathbb{N}$ and the space V, let

$$\mathrm{St}(N, V) := \{\boldsymbol{\phi} \in V \;:\; [\![\boldsymbol{\phi}, \boldsymbol{\phi}]\!]_H = \boldsymbol{I}_N\}$$

denote the infinite-dimensional *Stiefel manifold* of index N. Here, $\boldsymbol{I}_N$ is the identity matrix in $\mathbb{R}^{N\times N}$. We will see in Section 3 that $\mathrm{St}(N, V)$ admits a structure of an embedded submanifold of the Hilbert space V. Such a manifold was previously considered in [16, 33].

This paper is devoted to the abstract constrained energy minimization problem

$$\min_{\boldsymbol{\phi}\in\mathrm{St}(N,V)} \mathcal{E}(\boldsymbol{\phi}) \tag{2.6}$$

with the energy functional

$$\mathcal{E}(\boldsymbol{\phi}) := \frac{1}{2}\, a_0(\boldsymbol{\phi}, \boldsymbol{\phi}) + \frac{1}{2}\int_\Omega \Gamma(\rho(\boldsymbol{\phi}))\,\mathrm{d}r, \qquad \Gamma(\rho) = \int_0^\rho \gamma(t)\,\mathrm{d}t. \tag{2.7}$$

Throughout the paper, we make the (physically meaningful) assumption that $\mathcal{E}$ is orthogonally invariant in the sense that $\mathcal{E}(\boldsymbol{\phi}\,Q) = \mathcal{E}(\boldsymbol{\phi})$ for any orthogonal matrix $Q \in \mathbb{R}^{N\times N}$. This means that the energy depends only on

the space spanned by the components of $\boldsymbol{\phi}$ and not on a particular choice of $\boldsymbol{\phi}$. This condition is fulfilled in the applications we are interested in, see Section 5.

We are seeking critical points of the energy $\mathcal{E}$ which represent low-energy states. The state of minimal energy, which is called the *ground state*, is of particular interest. Critical points of the energy subject to the constraint are characterized by a coupled system of nonlinear eigenvector problems associated with the bilinear form $a_{\boldsymbol{\phi}}$ introduced in (2.3). The connection follows from the observation that the directional derivative $\mathrm{D}\mathcal{E}(\boldsymbol{\phi})[\boldsymbol{v}]$ of $\mathcal{E}$ at $\boldsymbol{\phi}$ along $\boldsymbol{v}$ is given by

$$\mathrm{D}\mathcal{E}(\boldsymbol{\phi})[\boldsymbol{v}] = a_{\boldsymbol{\phi}}(\boldsymbol{\phi}, \boldsymbol{v}) \qquad \text{for all } \boldsymbol{v} \in V. \tag{2.8}$$

The variational formulation of the nonlinear eigenvector problem then reads: seek $\boldsymbol{\phi} \in \mathrm{St}(N, V)$ and N eigenvalues $\lambda_1, \dots, \lambda_N \in \mathbb{R}$ such that

$$\tilde{a}_{\boldsymbol{\phi}}(\phi_j, v_j) = \lambda_j \, (\phi_j, v_j)_{L^2(\Omega)} \qquad \text{for all } (v_1, \dots, v_N) \in V. \tag{2.9}$$

We emphasize that all these problems are coupled, since the bilinear form $\tilde{a}_{\boldsymbol{\phi}}$ contains the information on the entire N-frame $\boldsymbol{\phi}$.

3. Geometry of the infinite-dimensional Stiefel manifold

In this section, we investigate the geometric structure of the Stiefel manifold $\mathrm{St}(N, V)$. First, we state that $\mathrm{St}(N, V)$ is an embedded submanifold of the Hilbert space V. This result can be proved analogously to the finite-dimensional case of the Stiefel matrix manifold; see Section 3.3.2 of [2].

Proposition 3.1. *The Stiefel manifold $\mathrm{St}(N, V)$ is a closed embedded submanifold of the Hilbert space V. It has co-dimension $N(N+1)/2$.*

The *tangent space* of $\mathrm{St}(N, V)$ at $\boldsymbol{\phi} \in \mathrm{St}(N, V)$ is given by

$$T_{\boldsymbol{\phi}} \mathrm{St}(N, V) := \{\boldsymbol{\eta} \in V \;\; : \;\; [\![\boldsymbol{\eta}, \boldsymbol{\phi}]\!]_H + [\![\boldsymbol{\phi}, \boldsymbol{\eta}]\!]_H = \mathbf{0}_N\}.$$

Hence, $T_{\boldsymbol{\phi}} \mathrm{St}(N, V)$ contains all functions $\boldsymbol{\eta} \in V$ for which the matrix $[\![\boldsymbol{\eta}, \boldsymbol{\phi}]\!]_H$ is skew-symmetric.

3.1. Hilbert metric and normal space

The simplest Riemannian metric on the Stiefel manifold $\mathrm{St}(N, V)$ is the *Hilbert metric* g_H inherited from the ambient space $V \subset H$. It is given by

$$g_H(\boldsymbol{\eta}, \boldsymbol{\zeta}) = (\boldsymbol{\eta}, \boldsymbol{\zeta})_H = \operatorname{tr} [\![\boldsymbol{\eta}, \boldsymbol{\zeta}]\!]_H \qquad \text{for all } \boldsymbol{\eta}, \boldsymbol{\zeta} \in T_{\boldsymbol{\phi}} \mathrm{St}(N, V).$$

This metric turns $\mathrm{St}(N, V)$ into a Riemannian submanifold of V. The *normal space at $\boldsymbol{\phi} \in \mathrm{St}(N, V)$ with respect to g_H* is then defined as

$$(T_{\boldsymbol{\phi}} \mathrm{St}(N, V))_H^{\perp} = \{\boldsymbol{z} \in V \;\; : \;\; g_H(\boldsymbol{z}, \boldsymbol{\eta}) = 0 \;\; \text{for all } \boldsymbol{\eta} \in T_{\boldsymbol{\phi}} \mathrm{St}(N, V)\}.$$

The following proposition gives an explicit characterization of this space. Its proof is similar to the finite-dimensional setting, which can be found in Section 2.2.1 of [14].

Proposition 3.2. *The normal space $(T_{\boldsymbol{\phi}} \mathrm{St}(N, V))_H^{\perp}$ at $\boldsymbol{\phi} \in \mathrm{St}(N, V)$ is given by*

$$(T_{\boldsymbol{\phi}} \mathrm{St}(N, V))_H^{\perp} = \{\boldsymbol{\phi}\, S \in V \;\; : \;\; S \in \mathcal{S}_{\mathrm{sym}}(N)\}, \tag{3.1}$$

where $\mathcal{S}_{\mathrm{sym}}(N)$ denotes the set of all real symmetric $N \times N$ matrices.

We now introduce an H-orthonormal basis of $(T_{\boldsymbol{\phi}}\,\mathrm{St}(N,V))_H^\perp$. Let $S^{ij} \in \mathcal{S}_{\mathrm{sym}}(N)$ denote the (normalized) symmetric matrix which has a non-zero entry at positions (i,j) and (j,i) and a zero otherwise. More precisely, we have

$$\begin{aligned} S^{ii} &= e_i\,e_i^T, && 1 \le i \le N, \\ S^{ij} &= \tfrac{1}{\sqrt{2}}\big(e_i\,e_j^T + e_j\,e_i^T\big), && 1 \le i < j \le N, \end{aligned} \tag{3.2}$$

where e_j denotes the jth column of $\boldsymbol{I}_N$. Note that these matrices form a basis of $\mathcal{S}_{\mathrm{sym}}(N)$. For $1 \le i \le j \le N$, we define the functions $\boldsymbol{\phi}^{ij} := \boldsymbol{\phi}\,S^{ij} \in (T_{\boldsymbol{\phi}}\,\mathrm{St}(N,V))_H^\perp$. This means that

$$\begin{aligned} \boldsymbol{\phi}^{ii} &= (0,\dots,0,\phi_i,0,\dots,0), && 1 \le i \le N, \\ \boldsymbol{\phi}^{ij} &= \tfrac{1}{\sqrt{2}}\,(0,\dots,0,\phi_j,0,\dots,0,\phi_i,\dots,0), && 1 \le i < j \le N, \end{aligned} \tag{3.3}$$

where ϕ_j (the jth component of $\boldsymbol{\phi}$) is placed at the ith position and ϕ_i at the jth position. Properties of these functions are summarized in the following proposition.

Proposition 3.3. *Let $\boldsymbol{\phi} \in \mathrm{St}(N,V)$. Then the functions $\boldsymbol{\phi}^{ij}$, $1 \le i \le j \le N$, introduced in (3.3) form an H-orthonormal basis of $(T_{\boldsymbol{\phi}}\,\mathrm{St}(N,V))_H^\perp$.*

Proof. First, we show the H-orthonormality of the functions $\boldsymbol{\phi}^{ij}$. Since $\boldsymbol{\phi} \in \mathrm{St}(N,V)$, we obtain

$$\big(\boldsymbol{\phi}^{ii},\boldsymbol{\phi}^{\ell\ell}\big)_H = \sum_{m=1}^N \big(\phi_m^{ii},\phi_m^{\ell\ell}\big)_{L^2(\Omega)} = \sum_{m=1}^N \delta_{im}\delta_{\ell m}\,(\phi_i,\phi_\ell)_{L^2(\Omega)} = \delta_{i\ell}.$$

For $k < \ell$, we have

$$\big(\boldsymbol{\phi}^{ii},\boldsymbol{\phi}^{k\ell}\big)_H = \big(\phi_i,\phi_i^{k\ell}\big)_{L^2(\Omega)} = \tfrac{1}{\sqrt{2}}\,\delta_{ik}(\phi_i,\phi_\ell)_{L^2(\Omega)} + \tfrac{1}{\sqrt{2}}\,\delta_{i\ell}(\phi_i,\phi_k)_{L^2(\Omega)} = \sqrt{2}\,\delta_{ik}\delta_{i\ell} = 0.$$

Finally, for $i < j$ and $k < \ell$, which implies $\delta_{i\ell}\delta_{jk} = 0$, we derive

$$\big(\boldsymbol{\phi}^{ij},\boldsymbol{\phi}^{k\ell}\big)_H = \sum_{m=1}^N \big(\phi_m^{ij},\phi_m^{k\ell}\big)_{L^2(\Omega)} = \frac{1}{2}\,(\delta_{ik}\delta_{j\ell} + \delta_{j\ell}\delta_{ik}) = \delta_{ik}\delta_{j\ell}.$$

Obviously, the functions $\boldsymbol{\phi}^{ij}$, $1 \le i \le j \le N$, span $(T_{\boldsymbol{\phi}}\,\mathrm{St}(N,V))_H^\perp$ and, hence, they form an H-orthonormal basis of $(T_{\boldsymbol{\phi}}\,\mathrm{St}(N,V))_H^\perp$. □

3.2. The $a_{\boldsymbol{\phi}}$-metric, normal space, and $a_{\boldsymbol{\phi}}$-orthogonal projection

An alternative Riemannian metric on the Stiefel manifold $\mathrm{St}(N,V)$ can be defined by using the inner product $a_{\boldsymbol{\phi}}(\cdot,\cdot)$ introduced in (2.3) as

$$g_a(\boldsymbol{\eta},\boldsymbol{\zeta}) = a_{\boldsymbol{\phi}}(\boldsymbol{\eta},\boldsymbol{\zeta}) \qquad \text{for all } \boldsymbol{\eta},\boldsymbol{\zeta} \in T_{\boldsymbol{\phi}}\,\mathrm{St}(N,V).$$

Then the *normal space at $\boldsymbol{\phi} \in \mathrm{St}(N,V)$ with respect to g_a* is defined as

$$(T_{\boldsymbol{\phi}}\,\mathrm{St}(N,V))_a^\perp = \{\boldsymbol{z} \in V \;:\; g_a(\boldsymbol{z},\boldsymbol{\eta}) = 0 \;\text{ for all } \boldsymbol{\eta} \in T_{\boldsymbol{\phi}}\,\mathrm{St}(N,V)\}.$$

Our goal is now to construct a basis of $(T_{\boldsymbol{\phi}}\,\mathrm{St}(N,V))_a^\perp$. To this end, we introduce the functions $\boldsymbol{\psi}^{k\ell} \in V$ for $1 \le k \le \ell \le N$ as solutions to

$$a_{\boldsymbol{\phi}}\big(\boldsymbol{\psi}^{k\ell},\boldsymbol{\eta}\big) = 0 \qquad \text{for all } \boldsymbol{\eta} \in T_{\boldsymbol{\phi}}\,\mathrm{St}(N,V), \tag{3.4a}$$

$$\big(\boldsymbol{\psi}^{k\ell},\boldsymbol{\phi}^{ij}\big)_H = \delta_{ik}\delta_{j\ell} \qquad \text{for } 1 \le i \le j \le N, \tag{3.4b}$$

where $\boldsymbol{\phi}^{ij}$ are defined in (3.3). The following proposition establishes the well-posedness of these problems.

Proposition 3.4. *There exist unique functions $\boldsymbol{\psi}^{k\ell} \in V$, $1 \le k \le \ell \le N$, satisfying (3.4).*

Proof. Let the indices $1 \le k \le \ell \le N$ be arbitrary but fixed. We can write (3.4) as a saddle point problem. Hence, we seek for $\boldsymbol{\psi}^{k\ell} \in V$ and Lagrange multipliers $\mu^{ij} \in \mathbb{R}$, $1 \le i \le j \le N$, such that

$$\begin{aligned} a_{\boldsymbol{\phi}}(\boldsymbol{\psi}^{k\ell}, \boldsymbol{v}) + \sum_{i \le j} (\boldsymbol{\phi}^{ij}, \boldsymbol{v})_H \mu^{ij} &= 0 && \text{for all } \boldsymbol{v} \in V, \\ (\boldsymbol{\psi}^{k\ell}, \boldsymbol{\phi}^{ij})_H &= \delta_{ik}\delta_{j\ell} && \text{for } 1 \le i \le j \le N. \end{aligned}$$

By Assumption 2.2, the bilinear form $a_{\boldsymbol{\phi}}$ is coercive. Moreover, the number of constraints equals $N(N+1)/2$ and is, hence, finite. In this case, the corresponding inf-sup stability follows from the linear independence of the functions $\boldsymbol{\phi}^{ij}$. As a result, Chapter III.4 of [8] implies the existence of a unique solution $\boldsymbol{\psi}^{k\ell} \in V$. Note that $\boldsymbol{\psi}^{k\ell}$ satisfies (3.4a), since by Proposition 3.3, we have $(\boldsymbol{\phi}^{ij}, \boldsymbol{\eta})_H = 0$ for all $\boldsymbol{\eta} \in T_{\boldsymbol{\phi}} \operatorname{St}(N, V)$. □

Next, we characterize the normal space $(T_{\boldsymbol{\phi}} \operatorname{St}(N, V))_a^\perp$ by providing a basis of it.

Proposition 3.5. *Let $\boldsymbol{\phi} \in \operatorname{St}(N, V)$. Then the functions $\boldsymbol{\psi}^{k\ell} \in V$, $1 \le k \le \ell \le N$, satisfying (3.4) form a basis of the normal space $(T_{\boldsymbol{\phi}} \operatorname{St}(N, V))_a^\perp$.*

Proof. It follows from (3.4a) that $\boldsymbol{\psi}^{k\ell} \in (T_{\boldsymbol{\phi}} \operatorname{St}(N, V))_a^\perp$. Further, equation (3.4b) implies that these functions are linearly independent. Taking into account that $(T_{\boldsymbol{\phi}} \operatorname{St}(N, V))_a^\perp$ has dimension $N(N+1)/2$, we obtain the result. □

Any element $\boldsymbol{v} \in V$ can be uniquely decomposed as $\boldsymbol{v} = P_{\boldsymbol{\phi}}(\boldsymbol{v}) + P_{\boldsymbol{\phi}}^\perp(\boldsymbol{v})$, where $P_{\boldsymbol{\phi}}$ and $P_{\boldsymbol{\phi}}^\perp$ denote the $a_{\boldsymbol{\phi}}$-orthogonal projections onto $T_{\boldsymbol{\phi}} \operatorname{St}(N, V)$ and $(T_{\boldsymbol{\phi}} \operatorname{St}(N, V))_a^\perp$, respectively. The projection operator $P_{\boldsymbol{\phi}}$ satisfies the conditions $P_{\boldsymbol{\phi}} \circ P_{\boldsymbol{\phi}} = P_{\boldsymbol{\phi}}$ and

$$[\![P_{\boldsymbol{\phi}}(\boldsymbol{v}), \boldsymbol{\phi}]\!]_H + [\![\boldsymbol{\phi}, P_{\boldsymbol{\phi}}(\boldsymbol{v})]\!]_H = \mathbf{0}_N, \tag{3.5a}$$

$$a_{\boldsymbol{\phi}}(\boldsymbol{v} - P_{\boldsymbol{\phi}}(\boldsymbol{v}), \boldsymbol{\eta}) = 0 \qquad \text{for all } \boldsymbol{\eta} \in T_{\boldsymbol{\phi}} \operatorname{St}(N, V). \tag{3.5b}$$

Note that (3.5) implies that $\operatorname{range} P_{\boldsymbol{\phi}} = T_{\boldsymbol{\phi}} \operatorname{St}(N, V)$ and $\ker P_{\boldsymbol{\phi}} = (T_{\boldsymbol{\phi}} \operatorname{St}(N, V))_a^\perp$. For the construction of such an operator, we use the basis functions $\boldsymbol{\psi}^{k\ell}$. It turns out that for any $\boldsymbol{v} \in V$, $P_{\boldsymbol{\phi}}(\boldsymbol{v})$ can be written as

$$\begin{aligned} P_{\boldsymbol{\phi}}(\boldsymbol{v}) &= \boldsymbol{v} - \sum_{k \le \ell} \big(\boldsymbol{v}, \boldsymbol{\phi}^{k\ell}\big)_H \boldsymbol{\psi}^{k\ell} \\ &= \boldsymbol{v} - \sum_{k=1}^{N} (v_k, \phi_k)_{L^2(\Omega)} \boldsymbol{\psi}^{kk} - \frac{1}{\sqrt{2}} \sum_{k<\ell} \Big[(v_k, \phi_\ell)_{L^2(\Omega)} + (v_\ell, \phi_k)_{L^2(\Omega)} \Big] \boldsymbol{\psi}^{k\ell}. \end{aligned} \tag{3.6}$$

The following result shows that this operator indeed satisfies the requested conditions and, hence, equals the $a_{\boldsymbol{\phi}}$-orthogonal projection onto $T_{\boldsymbol{\phi}} \operatorname{St}(N, V)$.

Proposition 3.6. *For $\boldsymbol{\phi} \in \operatorname{St}(N, V)$, the operator $P_{\boldsymbol{\phi}}$ from (3.6) is the $a_{\boldsymbol{\phi}}$-orthogonal projection onto $T_{\boldsymbol{\phi}} \operatorname{St}(N, V)$.*

Proof. First, we emphasize that $P_{\boldsymbol{\phi}}$ in (3.6) is a projection, since by Proposition 3.3 all summands $(\boldsymbol{v}, \boldsymbol{\phi}^{k\ell})_H$ vanish if $\boldsymbol{v}$ is already an element of $T_{\boldsymbol{\phi}} \operatorname{St}(N, V)$.

Next, we verify condition (3.5a), which means that $P_{\boldsymbol{\phi}}$ maps V into $T_{\boldsymbol{\phi}} \operatorname{St}(N, V)$. Note that for $k < \ell$, we obtain from (3.4b) that

$$\big(\psi_i^{k\ell}, \phi_i\big)_{L^2(\Omega)} = \big(\boldsymbol{\psi}^{k\ell}, \boldsymbol{\phi}^{ii}\big)_H = \delta_{ik}\delta_{i\ell} = 0, \qquad i = 1, \dots, N.$$

This implies $\big([\![P_{\boldsymbol{\phi}}(\boldsymbol{v}),\boldsymbol{\phi}]\!]_H + [\![\boldsymbol{\phi},P_{\boldsymbol{\phi}}(\boldsymbol{v})]\!]_H\big)_{i,i} = 2\,\big([\![P_{\boldsymbol{\phi}}(\boldsymbol{v}),\boldsymbol{\phi}]\!]_H\big)_{i,i} = 0$ for all $\boldsymbol{v} \in V$. Further, for $i \neq j$, we observe that

$$\begin{aligned}
\big([\![P_{\boldsymbol{\phi}}(\boldsymbol{v}),\boldsymbol{\phi}]\!]_H + [\![\boldsymbol{\phi},P_{\boldsymbol{\phi}}(\boldsymbol{v})]\!]_H\big)_{i,j} &= \big((P_{\boldsymbol{\phi}}(\boldsymbol{v}))_i,\phi_j\big)_{L^2(\Omega)} + \Big((P_{\boldsymbol{\phi}}(\boldsymbol{v}))_j,\phi_i\Big)_{L^2(\Omega)} \\
&= (v_i,\phi_j)_{L^2(\Omega)} + (v_j,\phi_i)_{L^2(\Omega)} \\
&\quad - \frac{1}{\sqrt{2}}\sum_{k<\ell}\Big[(v_k,\phi_\ell)_{L^2(\Omega)} + (v_\ell,\phi_k)_{L^2(\Omega)}\Big]\Big[\big(\psi_i^{k\ell},\phi_j\big)_{L^2(\Omega)} + \big(\psi_j^{k\ell},\phi_i\big)_{L^2(\Omega)}\Big] \\
&= (v_i,\phi_j)_{L^2(\Omega)} + (v_j,\phi_i)_{L^2(\Omega)} - \sum_{k<\ell}\Big[(v_k,\phi_\ell)_{L^2(\Omega)} + (v_\ell,\phi_k)_{L^2(\Omega)}\Big]\delta_{ik}\delta_{j\ell} = 0.
\end{aligned}$$

Finally, we show the $a_{\boldsymbol{\phi}}$-orthogonality property (3.5b). Indeed, for any $\boldsymbol{\eta} \in T_{\boldsymbol{\phi}}\,\mathrm{St}(N,V)$, equations (3.6) and (3.4a) yield

$$a_{\boldsymbol{\phi}}(\boldsymbol{v} - P_{\boldsymbol{\phi}}(\boldsymbol{v}),\boldsymbol{\eta}) = \sum_{k\leq\ell}\big(\boldsymbol{v},\boldsymbol{\phi}^{k\ell}\big)_H\, a_{\boldsymbol{\phi}}\big(\boldsymbol{\psi}^{k\ell},\boldsymbol{\eta}\big) = 0.$$

Thus, $P_{\boldsymbol{\phi}}$ is the $a_{\boldsymbol{\phi}}$-orthogonal projection onto $T_{\boldsymbol{\phi}}\,\mathrm{St}(N,V)$. □

For the Riemannian gradient descent method, which will be introduced in Section 4, we are especially interested in the projection operator $P_{\boldsymbol{\phi}}$ applied to $\boldsymbol{\phi} \in \mathrm{St}(N,V)$. In this case, we get

$$P_{\boldsymbol{\phi}}(\boldsymbol{\phi}) = \boldsymbol{\phi} - \sum_{k=1}^{N}(\phi_k,\phi_k)_{L^2(\Omega)}\,\boldsymbol{\psi}^{kk} = \boldsymbol{\phi} - \sum_{k=1}^{N}\boldsymbol{\psi}^{kk}.$$

Hence, for the computation of $P_{\boldsymbol{\phi}}(\boldsymbol{\phi})$, one only needs the sum $\boldsymbol{\psi} := \sum_{k=1}^{N}\boldsymbol{\psi}^{kk}$ of the functions $\boldsymbol{\psi}^{kk} \in V$, $k = 1,\dots,N$. It follows from (3.4) that this sum is uniquely defined by the equations

$$a_{\boldsymbol{\phi}}(\boldsymbol{\psi},\boldsymbol{\eta}) = 0 \qquad \text{for all } \boldsymbol{\eta} \in T_{\boldsymbol{\phi}}\,\mathrm{St}(N,V), \tag{3.7a}$$

$$(\boldsymbol{\psi},\boldsymbol{\phi}S^{ij})_H = \delta_{ij} \qquad \text{for } 1 \leq i \leq j \leq N. \tag{3.7b}$$

The following proposition provides an explicit expression for the solution $\boldsymbol{\psi}$.

Proposition 3.7. *Let $\boldsymbol{\phi} \in \mathrm{St}(N,V)$. The unique solution of system (3.7) is given by*

$$\boldsymbol{\psi} = \mathcal{A}_{\boldsymbol{\phi}}^{-1}\boldsymbol{\phi}\,\Big[\!\Big[\boldsymbol{\phi},\mathcal{A}_{\boldsymbol{\phi}}^{-1}\boldsymbol{\phi}\Big]\!\Big]_H^{-1}. \tag{3.8}$$

Proof. System (3.7) is equivalent to the saddle point problem

$$a_{\boldsymbol{\phi}}(\boldsymbol{\psi},\boldsymbol{v}) + \sum_{i\leq j}\big(\boldsymbol{\phi}S^{ij},\boldsymbol{v}\big)_H\mu^{ij} = 0 \qquad \text{for all } \boldsymbol{v} \in V, \tag{3.9a}$$

$$\big(\boldsymbol{\psi},\boldsymbol{\phi}S^{ij}\big)_H = \delta_{ij} \qquad \text{for } 1 \leq i \leq j \leq N \tag{3.9b}$$

for $\boldsymbol{\psi} \in V$ and the Lagrange multipliers $\mu^{ij} \in \mathbb{R}$. Using the special structure of the matrices S^{ij} in (3.2), the constraint conditions (3.9b) can be written as $\mathrm{sym}([\![\boldsymbol{\psi},\boldsymbol{\phi}]\!]_H) = \boldsymbol{I}_N$, where $\mathrm{sym}(A) = \frac{1}{2}(A + A^T)$ denotes the symmetric part of a matrix $A \in \mathbb{R}^{N\times N}$. Further, we obtain

$$\sum_{i\leq j}\big(\boldsymbol{\phi}S^{ij},\boldsymbol{v}\big)_H\mu^{ij} = \left(\boldsymbol{\phi}\sum_{i\leq j}S^{ij}\mu^{ij},\boldsymbol{v}\right)_H = (\boldsymbol{\phi}\,S,\boldsymbol{v})_H$$

with the symmetric matrix $S = \sum_{i \le j} S^{ij} \mu^{ij}$. As a result, system (3.9) takes the form

$$a_{\boldsymbol{\phi}}(\boldsymbol{\psi}, \boldsymbol{v}) + (\boldsymbol{\phi}\, S, \boldsymbol{v})_H = 0 \qquad \text{for all } \boldsymbol{v} \in V, \tag{3.10a}$$

$$\operatorname{sym}([\![\boldsymbol{\psi}, \boldsymbol{\phi}]\!]_H) \qquad = \boldsymbol{I}_N. \tag{3.10b}$$

Using (2.5), we derive from (3.10a) that

$$0 = a_{\boldsymbol{\phi}}(\boldsymbol{\psi}, \boldsymbol{v}) + a_{\boldsymbol{\phi}}\Big(\mathcal{A}_{\boldsymbol{\phi}}^{-1} \boldsymbol{\phi}\, S, \boldsymbol{v}\Big) = a_{\boldsymbol{\phi}}\Big(\boldsymbol{\psi} + \mathcal{A}_{\boldsymbol{\phi}}^{-1} \boldsymbol{\phi}\, S, \boldsymbol{v}\Big) \qquad \text{for all } \boldsymbol{v} \in V$$

and, hence, $\boldsymbol{\psi} = -\mathcal{A}_{\boldsymbol{\phi}}^{-1} \boldsymbol{\phi}\, S$. Substituting this function into (3.10b) yields the Lyapunov equation

$$\Big[\!\Big[\boldsymbol{\phi}, \mathcal{A}_{\boldsymbol{\phi}}^{-1} \boldsymbol{\phi}\Big]\!\Big]_H S + S \Big[\!\Big[\boldsymbol{\phi}, \mathcal{A}_{\boldsymbol{\phi}}^{-1} \boldsymbol{\phi}\Big]\!\Big]_H = -2\boldsymbol{I}_N \tag{3.11}$$

for S. By Proposition 2.3, the matrix $\Big[\!\Big[\boldsymbol{\phi}, \mathcal{A}_{\boldsymbol{\phi}}^{-1} \boldsymbol{\phi}\Big]\!\Big]_H$ is symmetric positive definite. In this case, the Lyapunov equation (3.11) has a unique symmetric solution ([24], Thm. 12.3.2) given by $S = -\Big[\!\Big[\boldsymbol{\phi}, \mathcal{A}_{\boldsymbol{\phi}}^{-1} \boldsymbol{\phi}\Big]\!\Big]_H^{-1}$. This finally gives the expression (3.8). □

3.3. Retractions

Next, we introduce the concept of retractions on the Stiefel manifold $\operatorname{St}(N, V)$. Retractions provide a useful tool in Riemannian optimization which allows us to keep the iteration points on the manifold.

Definition 3.8 (Retraction)**.** Let $T \operatorname{St}(N, V)$ be the tangent bundle of $\operatorname{St}(N, V)$. A smooth map $\mathcal{R} \colon T \operatorname{St}(N, V) \to \operatorname{St}(N, V)$ is called a *retraction on* $\operatorname{St}(N, V)$ if for all $\boldsymbol{\phi} \in \operatorname{St}(N, V)$, the restriction $\mathcal{R}_{\boldsymbol{\phi}} = \mathcal{R}\big|_{T_{\boldsymbol{\phi}} \operatorname{St}(N,V)}$ on $T_{\boldsymbol{\phi}} \operatorname{St}(N, V)$ has the following properties:

(a) $\mathcal{R}_{\boldsymbol{\phi}}(\mathbf{0}_{\boldsymbol{\phi}}) = \mathcal{R}(\boldsymbol{\phi}, \mathbf{0}_{\boldsymbol{\phi}}) = \boldsymbol{\phi}$, where $\mathbf{0}_{\boldsymbol{\phi}}$ denotes the zero element of $T_{\boldsymbol{\phi}} \operatorname{St}(N, V)$,

(b) $\frac{\mathrm{d}}{\mathrm{d}t} \mathcal{R}_{\boldsymbol{\phi}}(t\boldsymbol{\eta})\big|_{t=0} = \boldsymbol{\eta}$ for all $\boldsymbol{\eta} \in T_{\boldsymbol{\phi}} \operatorname{St}(N, V)$.

In Example 4.1.3 of [2] and [1,14,22,30], several retractions on the (generalized) Stiefel matrix manifold have been introduced and compared with respect to computational cost and accuracy. Here, we extend some of the decomposition-based retractions to the manifold $\operatorname{St}(N, V)$.

3.3.1. The projective retraction

First, we introduce a retraction based on the polar decomposition and show that it provides a projection onto $\operatorname{St}(N, V)$.

Similarly to the matrix case, *e.g.*, Section 9.4.3 of [15], we define the *polar decomposition* of $\boldsymbol{v} \in V$ as $\boldsymbol{v} = \boldsymbol{u}S$, where $\boldsymbol{u} \in \operatorname{St}(N, V)$ and $S \in \mathbb{R}^{N \times N}$ is symmetric positive semidefinite. Such a decomposition always exists. If the components of $\boldsymbol{v}$ are linearly independent, then the matrix $[\![\boldsymbol{v}, \boldsymbol{v}]\!]_H$ is positive definite. In this case, $S = [\![\boldsymbol{v}, \boldsymbol{v}]\!]_H^{1/2}$ is positive definite and the factor $\boldsymbol{u} = \boldsymbol{v}\, [\![\boldsymbol{v}, \boldsymbol{v}]\!]_H^{-1/2}$ is unique.

For any $(\boldsymbol{\phi}, \boldsymbol{\eta}) \in T \operatorname{St}(N, V)$, *i.e.*, $\boldsymbol{\eta} \in T_{\boldsymbol{\phi}} \operatorname{St}(N, V)$, the components of $\boldsymbol{\phi} + \boldsymbol{\eta}$ are linearly independent, since the matrix

$$[\![\boldsymbol{\phi} + \boldsymbol{\eta}, \boldsymbol{\phi} + \boldsymbol{\eta}]\!]_H = [\![\boldsymbol{\phi}, \boldsymbol{\phi}]\!]_H + [\![\boldsymbol{\phi}, \boldsymbol{\eta}]\!]_H + [\![\boldsymbol{\eta}, \boldsymbol{\phi}]\!]_H + [\![\boldsymbol{\eta}, \boldsymbol{\eta}]\!]_H = \boldsymbol{I}_N + [\![\boldsymbol{\eta}, \boldsymbol{\eta}]\!]_H \tag{3.12}$$

is positive definite. Then we can use the polar decomposition of $\boldsymbol{\phi} + \boldsymbol{\eta}$ to define a retraction on $\operatorname{St}(N, V)$.

Proposition 3.9. *For $(\boldsymbol{\phi}, \boldsymbol{\eta}) \in T \operatorname{St}(N, V)$, the map*

$$\mathcal{R}(\boldsymbol{\phi}, \boldsymbol{\eta}) := (\boldsymbol{\phi} + \boldsymbol{\eta})(\boldsymbol{I}_N + [\![\boldsymbol{\eta}, \boldsymbol{\eta}]\!]_H)^{-1/2} \tag{3.13}$$

is a retraction on $\operatorname{St}(N, V)$.

Proof. Let $(\boldsymbol{\phi}, \boldsymbol{\eta}) \in T\,\mathrm{St}(N,V)$. First, we verify that $\mathcal{R}(\boldsymbol{\phi}, \boldsymbol{\eta})$ belongs to $\mathrm{St}(N,V)$. Using Lemma 2.1 and (3.12), we obtain

$$\begin{aligned}\llbracket \mathcal{R}(\boldsymbol{\phi},\boldsymbol{\eta}), \mathcal{R}(\boldsymbol{\phi},\boldsymbol{\eta}) \rrbracket_H &= \Big\llbracket (\boldsymbol{\phi}+\boldsymbol{\eta})(\boldsymbol{I}_N + \llbracket \boldsymbol{\eta},\boldsymbol{\eta}\rrbracket_H)^{-1/2}, (\boldsymbol{\phi}+\boldsymbol{\eta})(\boldsymbol{I}_N + \llbracket \boldsymbol{\eta},\boldsymbol{\eta}\rrbracket_H)^{-1/2} \Big\rrbracket_H \\ &= (\boldsymbol{I}_N + \llbracket \boldsymbol{\eta},\boldsymbol{\eta}\rrbracket_H)^{-1/2} (\boldsymbol{I}_N + \llbracket \boldsymbol{\eta},\boldsymbol{\eta}\rrbracket_H)(\boldsymbol{I}_N + \llbracket \boldsymbol{\eta},\boldsymbol{\eta}\rrbracket_H)^{-1/2} = \boldsymbol{I}_N,\end{aligned}$$

and, hence, $\mathcal{R}(\boldsymbol{\phi}, \boldsymbol{\eta}) \in \mathrm{St}(N,V)$. Furthermore, we have $\mathcal{R}_{\boldsymbol{\phi}}(\mathbf{0}_{\boldsymbol{\phi}}) = \boldsymbol{\phi}$ and

$$\begin{aligned}\tfrac{\mathrm{d}}{\mathrm{d}t}\mathcal{R}_{\boldsymbol{\phi}}(t\boldsymbol{\eta})\Big|_{t=0} &= \tfrac{\mathrm{d}}{\mathrm{d}t}(\boldsymbol{\phi}+t\boldsymbol{\eta})\big(\boldsymbol{I}_N + t^2\llbracket \boldsymbol{\eta},\boldsymbol{\eta}\rrbracket_H\big)^{-1/2}\Big|_{t=0} \\ &= -t\,(\boldsymbol{\phi}+t\boldsymbol{\eta})\,\llbracket \boldsymbol{\eta},\boldsymbol{\eta}\rrbracket_H \big(\boldsymbol{I}_N + t^2\llbracket \boldsymbol{\eta},\boldsymbol{\eta}\rrbracket_H\big)^{-3/2} + \boldsymbol{\eta}\,\big(\boldsymbol{I}_N + t^2\llbracket \boldsymbol{\eta},\boldsymbol{\eta}\rrbracket_H\big)^{-1/2}\Big|_{t=0} \\ &= \boldsymbol{\eta}.\end{aligned}$$

This shows that $\mathcal{R}$ defined in (3.13) is the retraction on $\mathrm{St}(N,V)$. □

The evaluation of the retraction in (3.13) involves the computation of the outer product $\llbracket \boldsymbol{\eta},\boldsymbol{\eta}\rrbracket_H$ and the eigenvalue decomposition

$$\boldsymbol{I}_N + \llbracket \boldsymbol{\eta},\boldsymbol{\eta}\rrbracket_H = QDQ^T, \tag{3.14}$$

where $Q \in \mathbb{R}^{N\times N}$ is orthogonal and $D = \mathrm{diag}(d_1,\ldots,d_N)$ with $d_j > 0$ for $j = 1,\ldots,N$. With this, we obtain $\mathcal{R}(\boldsymbol{\phi},\boldsymbol{\eta}) = (\boldsymbol{\phi}+\boldsymbol{\eta})\,QD^{-1/2}Q^T$.

Remark 3.10. For stability reasons, we recommend to use $\llbracket \boldsymbol{\phi}+\boldsymbol{\eta}, \boldsymbol{\phi}+\boldsymbol{\eta}\rrbracket_H$ instead of $\boldsymbol{I}_N + \llbracket \boldsymbol{\eta},\boldsymbol{\eta}\rrbracket_H$ in (3.14). A similar suggestion for the generalized Stiefel matrix manifold can be found in [30]. Note that, due to (3.12), both expressions are equivalent if $\boldsymbol{\phi} \in \mathrm{St}(N,V)$ and $\boldsymbol{\eta} \in T_{\boldsymbol{\phi}}\,\mathrm{St}(N,V)$.

The polar decomposition based retraction (3.13) can be viewed as a projective retraction, since it satisfies

$$\mathcal{R}(\boldsymbol{\phi},\boldsymbol{\eta}) = \operatorname*{arg\,min}_{\boldsymbol{\xi}\in\mathrm{St}(N,V)} \|\boldsymbol{\xi} - (\boldsymbol{\phi}+\boldsymbol{\eta})\|_H^2. \tag{3.15}$$

To prove this, we first observe that for all $(\boldsymbol{\phi},\boldsymbol{\eta}) \in T\,\mathrm{St}(N,V)$, $\boldsymbol{\phi}+\boldsymbol{\eta}$ can be represented as

$$\boldsymbol{\phi}+\boldsymbol{\eta} = \boldsymbol{u}D^{1/2}Q^T, \tag{3.16}$$

where $\boldsymbol{u} \in \mathrm{St}(N,V)$ and D, Q are as in (3.14). This decomposition is an extension of the singular value decomposition known for matrices, *e.g.*, Section 2.4 of [15] to the elements of V. For any $\boldsymbol{\xi} \in \mathrm{St}(N,V)$, we have

$$\|\boldsymbol{\xi} - (\boldsymbol{\phi}+\boldsymbol{\eta})\|_H^2 = \|\boldsymbol{\xi}\|_H^2 - 2\,(\boldsymbol{\xi}, \boldsymbol{\phi}+\boldsymbol{\eta})_H + \|\boldsymbol{\phi}+\boldsymbol{\eta}\|_H^2 = N^2 - 2\,(\boldsymbol{\xi}, \boldsymbol{\phi}+\boldsymbol{\eta})_H + \mathrm{tr}\,D$$

with

$$\begin{aligned}(\boldsymbol{\xi}, \boldsymbol{\phi}+\boldsymbol{\eta})_H &= \mathrm{tr}\left(\left\llbracket \boldsymbol{\xi}, \boldsymbol{u}\,D^{1/2}Q^T \right\rrbracket_H\right) = \mathrm{tr}\left(\llbracket \boldsymbol{\xi},\boldsymbol{u}\rrbracket_H D^{1/2}\right) \\ &= \sum_{i=1}^N (\xi_i, u_i)_{L^2(\Omega)}\sqrt{d_i} \le \sum_{i=1}^N \|\xi_i\|_{L^2(\Omega)}\|u_i\|_{L^2(\Omega)}\sqrt{d_i} = \mathrm{tr}\,D^{1/2}.\end{aligned}$$

For $\boldsymbol{\xi} = \boldsymbol{u}\,Q^T \in \mathrm{St}(N,V)$, the equality

$$(\boldsymbol{\xi}, \boldsymbol{\phi}+\boldsymbol{\eta})_H = \mathrm{tr}\left\llbracket \boldsymbol{u}\,Q^T, \boldsymbol{u}\,D^{1/2}Q^T \right\rrbracket_H = \mathrm{tr}\,D^{1/2}$$

holds, *i.e.*, $\boldsymbol{\xi} = \boldsymbol{u}\,Q^T$ solves (3.15). Thus, $\mathcal{R}(\boldsymbol{\phi},\boldsymbol{\eta}) = (\boldsymbol{\phi}+\boldsymbol{\eta})QD^{-1/2}Q^T = \boldsymbol{u}\,Q^T$ is a projection onto $\mathrm{St}(N,V)$.

The following proposition shows that the retraction (3.13) is second-order bounded.

Proposition 3.11. *The retraction $\mathcal{R}$ in (3.13) satisfies*

$$\|\mathcal{R}(\boldsymbol{\phi}, t\boldsymbol{\eta}) - (\boldsymbol{\phi} + t\boldsymbol{\eta})\|_{a_{\boldsymbol{\phi}}} \leq t^2 \, \|\boldsymbol{\phi} + t\boldsymbol{\eta}\|_{a_{\boldsymbol{\phi}}} \|\boldsymbol{\eta}\|_H^2.$$

Proof. The proof is given in Appendix A.1. □

3.3.2. The qR-based retraction

An alternative retraction on $\mathrm{St}(N, V)$ can be defined by using the orthonormalization with respect to the inner product $(\cdot, \cdot)_H$. First, we observe that for any $\boldsymbol{v} = (v_1, \ldots, v_N) \in V$ with linearly independent components, there exist $\boldsymbol{q} \in \mathrm{St}(N, V)$ and an upper triangular matrix $R \in \mathbb{R}^{N \times N}$ with strictly positive diagonal elements such that $\boldsymbol{v} = \boldsymbol{q}R$. The existence of such a decomposition, called *qR decomposition*, can be proved constructively by using the Gram–Schmidt orthonormalization procedure

$$\begin{aligned} \tilde{q}_1 &:= v_1, & q_1 &:= \frac{\tilde{q}_1}{\|\tilde{q}_1\|_{L^2(\Omega)}}, \\ \tilde{q}_j &:= v_j - \sum_{i=1}^{j-1} (v_j, q_i)_{L^2(\Omega)} \, q_i, & q_j &:= \frac{\tilde{q}_j}{\|\tilde{q}_j\|_{L^2(\Omega)}}, \qquad j = 2, \ldots, N. \end{aligned} \tag{3.17}$$

With this, we obtain $\boldsymbol{q} = (q_1, \ldots, q_N) \in \mathrm{St}(N, V)$ and

$$R = \begin{bmatrix} (v_1, q_1)_{L^2(\Omega)} & (v_2, q_1)_{L^2(\Omega)} & \cdots & \cdots & (v_N, q_1)_{L^2(\Omega)} \\ 0 & (v_2, q_2)_{L^2(\Omega)} & \cdots & \cdots & (v_N, q_2)_{L^2(\Omega)} \\ 0 & 0 & \ddots & & \vdots \\ \vdots & & \ddots & \ddots & \vdots \\ 0 & 0 & \cdots & 0 & (v_N, q_N)_{L^2(\Omega)} \end{bmatrix}.$$

Note that the matrix R has positive diagonal elements $(v_j, q_j)_{L^2(\Omega)} = \|\tilde{q}_j\|_{L^2(\Omega)}$. This property of R guarantees the uniqueness of the qR decomposition. Let $\mathrm{qf}(\boldsymbol{v})$ denote the factor $\boldsymbol{q}$ in $\boldsymbol{v} = \boldsymbol{q}R$. This allows us to define a qR-based retraction on the Stiefel manifold $\mathrm{St}(N, V)$.

Proposition 3.12. *For $(\boldsymbol{\phi}, \boldsymbol{\eta}) \in T\,\mathrm{St}(N, V)$, the map*

$$\mathcal{R}(\boldsymbol{\phi}, \boldsymbol{\eta}) := \mathrm{qf}(\boldsymbol{\phi} + \boldsymbol{\eta}) \tag{3.18}$$

is a retraction on $\mathrm{St}(N, V)$.

Proof. Obviously, $\mathcal{R}$ in (3.18) is well defined on $T\,\mathrm{St}(N, V)$. Further, by definition, we have $\mathcal{R}(\boldsymbol{\phi}, \boldsymbol{\eta}) \in \mathrm{St}(N, V)$ for all $(\boldsymbol{\phi}, \boldsymbol{\eta}) \in T\,\mathrm{St}(N, V)$ and $\mathcal{R}_{\boldsymbol{\phi}}(\mathbf{0}_{\boldsymbol{\phi}}) = \boldsymbol{\phi}$.

In order to prove the second property in Definition 3.8, we follow the lines of Example 8.1.5 from [2]. For any $(\boldsymbol{\phi}, \boldsymbol{\eta}) \in T\,\mathrm{St}(N, V)$, we consider a curve $\boldsymbol{\varphi}(t) = \boldsymbol{\phi} + t\boldsymbol{\eta}$. Let $\boldsymbol{\varphi}(t) = \boldsymbol{q}(t)R(t)$ be the qR decomposition of $\boldsymbol{\varphi}(t)$. Then, using the product rule, we have

$$\dot{\boldsymbol{\varphi}}(t) = \dot{\boldsymbol{q}}(t)R(t) + \boldsymbol{q}(t)\dot{R}(t), \tag{3.19}$$

where $\dot{\boldsymbol{\varphi}}(t) = \frac{\mathrm{d}}{\mathrm{d}t}\boldsymbol{\varphi}(t)$ and similar for $\boldsymbol{q}(t)$ and $R(t)$. For the sake of brevity, we omit the argument t in what follows. Computing the outer product of $\boldsymbol{q}$ and $\dot{\boldsymbol{\varphi}}$, we obtain

$$[\![\boldsymbol{q}, \dot{\boldsymbol{\varphi}}]\!]_H = [\![\boldsymbol{q}, \dot{\boldsymbol{q}}]\!]_H R + [\![\boldsymbol{q}, \boldsymbol{q}]\!]_H \dot{R} = [\![\boldsymbol{q}, \dot{\boldsymbol{q}}]\!]_H R + \dot{R}. \tag{3.20}$$

Multiplication of (3.20) by R^{-1} from the right yields

$$[\![\boldsymbol{q}, \dot{\boldsymbol{\varphi}}]\!]_H R^{-1} = [\![\boldsymbol{q}, \dot{\boldsymbol{q}}]\!]_H + \dot{R}R^{-1},$$

Algorithm 1. Modified Gram–Schmidt procedure.

Input: $\boldsymbol{v} = (v_1, \dots, v_N) \in V$
for $i = 1, \dots, N$ **do**
  $r_{ii} = \|v_i\|_{L^2(\Omega)}$
  $q_i = v_i / r_{ii}$
  for $j = i+1, \dots, N$ **do**
    $r_{ij} = (v_j, q_i)_{L^2(\Omega)}$
    $v_j = v_j - r_{ij} q_i$
Output: $\boldsymbol{q} = (q_1, \dots, q_N) \in \mathrm{St}(N, V)$ and $R = [r_{ij}] \in \mathbb{R}^{N \times N}$ such that $\boldsymbol{v} = \boldsymbol{q}R$

where $[\![\boldsymbol{q}, \dot{\boldsymbol{q}}]\!]_H$ is skew-symmetric and $\dot{R}R^{-1}$ is upper triangular. Since $M := [\![\boldsymbol{q}, \dot{\boldsymbol{\varphi}}]\!]_H R^{-1}$ can uniquely be represented as $M = \varrho_{\mathrm{skew}}(M) + \varrho_{\mathrm{up}}(M)$, where $\varrho_{\mathrm{skew}}(M)$ is skew-symmetric and $\varrho_{\mathrm{up}}(M)$ is upper triangular, we obtain

$$\varrho_{\mathrm{skew}}\big([\![\boldsymbol{q}, \dot{\boldsymbol{\varphi}}]\!]_H R^{-1}\big) = [\![\boldsymbol{q}, \dot{\boldsymbol{q}}]\!]_H, \qquad \varrho_{\mathrm{up}}\big([\![\boldsymbol{q}, \dot{\boldsymbol{\varphi}}]\!]_H R^{-1}\big) = \dot{R}R^{-1}.$$

Further, multiplying (3.20) by $\boldsymbol{q}$ from the left and subtracting the resulting equation from (3.19), we find

$$\dot{\boldsymbol{\varphi}} - \boldsymbol{q}\,[\![\boldsymbol{q}, \dot{\boldsymbol{\varphi}}]\!]_H = \dot{\boldsymbol{q}}R - \boldsymbol{q}\,\varrho_{\mathrm{skew}}\big([\![\boldsymbol{q}, \dot{\boldsymbol{\varphi}}]\!]_H R^{-1}\big)R,$$

which implies

$$\dot{\boldsymbol{q}} = (\dot{\boldsymbol{\varphi}} - \boldsymbol{q}\,[\![\boldsymbol{q}, \dot{\boldsymbol{\varphi}}]\!]_H)R^{-1} + \boldsymbol{q}\,\varrho_{\mathrm{skew}}\big([\![\boldsymbol{q}, \dot{\boldsymbol{\varphi}}]\!]_H R^{-1}\big).$$

Taking into account that $\dot{\boldsymbol{\varphi}}(0) = \boldsymbol{\eta}$, $\boldsymbol{\varphi}(0) = \boldsymbol{\phi} = \boldsymbol{q}(0)$, $R(0) = \boldsymbol{I}_N$, and that $[\![\boldsymbol{\phi}, \boldsymbol{\eta}]\!]_H$ is skew-symmetric, we finally obtain

$$\tfrac{\mathrm{d}}{\mathrm{d}t}\mathcal{R}_{\boldsymbol{\phi}}(t\boldsymbol{\eta})\Big|_{t=0} = \tfrac{\mathrm{d}}{\mathrm{d}t}\,\boldsymbol{q}(t)\Big|_{t=0} = \boldsymbol{\eta} - \boldsymbol{\phi}\,[\![\boldsymbol{\phi}, \boldsymbol{\eta}]\!]_H + \boldsymbol{\phi}\,[\![\boldsymbol{\phi}, \boldsymbol{\eta}]\!]_H = \boldsymbol{\eta},$$

which completes the proof. □

The qR-based retraction (3.18) can be computed by the modified Gram–Schmidt procedure as presented in Algorithm 1 which is more numerically stable than the Gram–Schmidt process (3.17). An alternative approach for evaluating (3.18) is based on computing the Cholesky factorization $[\![\boldsymbol{\phi} + \boldsymbol{\eta}, \boldsymbol{\phi} + \boldsymbol{\eta}]\!]_H = F^T F$ with an upper triangular matrix $F \in \mathbb{R}^{N \times N}$ and determining

$$\mathcal{R}(\boldsymbol{\phi}, \boldsymbol{\eta}) = (\boldsymbol{\phi} + \boldsymbol{\eta})\,F^{-1}. \tag{3.21}$$

It is an extension of the Cholesky-QR-based method on the generalized matrix Stiefel manifold presented in [30]. Compared to the polar decomposition based retraction (3.13), the computation of (3.21) has lower numerical complexity, especially for large N, since it requires the Cholesky factorization instead of the eigenvalue decomposition.

The following proposition establishes the second-order boundedness of the qR-based retraction (3.18).

Proposition 3.13. *The retraction $\mathcal{R}$ in* (3.18) *satisfies*

$$\|\mathcal{R}(\boldsymbol{\phi}, t\boldsymbol{\eta}) - (\boldsymbol{\phi} + t\boldsymbol{\eta})\|_{a_{\boldsymbol{\phi}}} \leq \frac{t^2}{\sqrt{2}}\,\|\boldsymbol{\phi} + t\boldsymbol{\eta}\|_{a_{\boldsymbol{\phi}}}\big(1 + t^2\|\boldsymbol{\eta}\|_H^2\big)^{1/2}\|\boldsymbol{\eta}\|_H^2.$$

Proof. The proof is given in Appendix A.2. □

4. Energy-adaptive Riemannian gradient descent method

The simplest approach to minimize the energy functional $\mathcal{E}$ over $\mathrm{St}(N,V)$ is the gradient descent method, which requires the Riemannian gradient of $\mathcal{E}$. For a smooth scalar field $\mathcal{E}$ on the Riemannian manifold $\mathrm{St}(N,V)$, the *Riemannian gradient* $\operatorname{grad}\mathcal{E}(\boldsymbol{\phi})$ of $\mathcal{E}$ at $\boldsymbol{\phi}\in\mathrm{St}(N,V)$ with respect to the metric g_a is defined as the unique element of the tangent space $T_{\boldsymbol{\phi}}\,\mathrm{St}(N,V)$ satisfying

$$g_a(\operatorname{grad}\mathcal{E}(\boldsymbol{\phi}),\boldsymbol{\eta}) = a_{\boldsymbol{\phi}}(\operatorname{grad}\mathcal{E}(\boldsymbol{\phi}),\boldsymbol{\eta}) = \mathrm{D}\mathcal{E}(\boldsymbol{\phi})[\boldsymbol{\eta}] \qquad \text{for all } \boldsymbol{\eta}\in T_{\boldsymbol{\phi}}\,\mathrm{St}(N,V).$$

Since $\mathrm{St}(N,V)$ is an embedded submanifold of V, we obtain the following expression for the Riemannian gradient.

Proposition 4.1. *The Riemannian gradient of the energy functional $\mathcal{E}\colon V\to\mathbb{R}$ from (2.7) at $\boldsymbol{\phi}\in\mathrm{St}(N,V)$ with respect to the metric g_a is given by*

$$\operatorname{grad}\mathcal{E}(\boldsymbol{\phi}) = P_{\boldsymbol{\phi}}(\boldsymbol{\phi}) = \boldsymbol{\phi} - \mathcal{A}_{\boldsymbol{\phi}}^{-1}\boldsymbol{\phi}\left[\!\left[\boldsymbol{\phi},\mathcal{A}_{\boldsymbol{\phi}}^{-1}\boldsymbol{\phi}\right]\!\right]_H^{-1}. \tag{4.1}$$

Proof. Using (2.8), we obtain

$$a_{\boldsymbol{\phi}}(\operatorname{grad}\mathcal{E}(\boldsymbol{\phi}),\boldsymbol{\eta}) = \mathrm{D}\mathcal{E}(\boldsymbol{\phi})[\boldsymbol{\eta}] = a_{\boldsymbol{\phi}}(\boldsymbol{\phi},\boldsymbol{\eta}) \qquad \text{for all } \boldsymbol{\eta}\in T_{\boldsymbol{\phi}}\,\mathrm{St}(N,V).$$

Hence, $a_{\boldsymbol{\phi}}(\operatorname{grad}\mathcal{E}(\boldsymbol{\phi})-\boldsymbol{\phi},\boldsymbol{\eta})=0$ for all $\boldsymbol{\eta}\in T_{\boldsymbol{\phi}}\,\mathrm{St}(N,V)$. This implies that $\operatorname{grad}\mathcal{E}(\boldsymbol{\phi})-\boldsymbol{\phi}$ belongs to the normal space $(T_{\boldsymbol{\phi}}\,\mathrm{St}(N,V))_a^{\perp}$ and, hence,

$$\operatorname{grad}\mathcal{E}(\boldsymbol{\phi}) = \boldsymbol{\phi} + P_{\boldsymbol{\phi}}^{\perp}(\operatorname{grad}\mathcal{E}(\boldsymbol{\phi})-\boldsymbol{\phi}) = \boldsymbol{\phi} - P_{\boldsymbol{\phi}}^{\perp}(\boldsymbol{\phi}) = P_{\boldsymbol{\phi}}(\boldsymbol{\phi}).$$

The second expression for $\operatorname{grad}\mathcal{E}(\boldsymbol{\phi})$ in (4.1) immediately follows from Proposition 3.7. □

Using the Riemannian gradient and any retraction $\mathcal{R}$ on $\mathrm{St}(N,V)$ from Section 3.3, the Riemannian gradient descent method for solving the minimization problem (2.6) can be formulated as follows: for given $\boldsymbol{\phi}^{(n)}\in\mathrm{St}(N,V)$, compute

$$\boldsymbol{\phi}^{(n+1)} = \mathcal{R}\Big(\boldsymbol{\phi}^{(n)},\tau_n\boldsymbol{\eta}^{(n)}\Big) \tag{4.2}$$

with the search direction $\boldsymbol{\eta}^{(n)} = -\operatorname{grad}\mathcal{E}\big(\boldsymbol{\phi}^{(n)}\big)$ and an appropriately chosen step size $\tau_n>0$.

Remark 4.2 (Connection to Sobolev gradient flows)**.** The presented minimization approach for solving the nonlinear eigenvector problem (2.9) is closely related to the Sobolev gradient flow algorithm studied in [18] for the Gross–Pitaevskii eigenvalue problem which, as will be shown in Section 5, fits in the given framework with $N=1$. For general problems with $N\geq 1$, let $\nabla\mathcal{E}(\boldsymbol{\phi})$ denote the Riesz representative of $\mathrm{D}\mathcal{E}(\boldsymbol{\phi})$ in the Hilbert space V with respect to the inner product $a_{\boldsymbol{\phi}}(\cdot,\cdot)$. The operator $\nabla\mathcal{E}\colon V\to V$ is called the *$a_{\boldsymbol{\phi}}$-Sobolev gradient of $\mathcal{E}$*. It follows from (2.8) that

$$a_{\boldsymbol{\phi}}(\nabla\mathcal{E}(\boldsymbol{\phi}),\boldsymbol{v}) = \mathrm{D}\mathcal{E}(\boldsymbol{\phi})[\boldsymbol{v}] = a_{\boldsymbol{\phi}}(\boldsymbol{\phi},\boldsymbol{v}) \qquad \text{for all } \boldsymbol{v}\in V$$

and, hence, $\nabla\mathcal{E}(\boldsymbol{\phi})=\boldsymbol{\phi}$. Given an initial guess $\boldsymbol{\phi}(0)\in\mathrm{St}(N,V)$, the corresponding dynamical system, also called the $a_{\boldsymbol{\phi}}$-Sobolev gradient flow, has the form

$$\dot{\boldsymbol{\phi}}(t) = -P_{\boldsymbol{\phi}(t)}(\nabla\mathcal{E}(\boldsymbol{\phi}(t))) = -P_{\boldsymbol{\phi}(t)}(\boldsymbol{\phi}(t)) = -\operatorname{grad}\mathcal{E}(\boldsymbol{\phi}(t)). \tag{4.3}$$

It can be easily seen that the solution of this system satisfies $\boldsymbol{\phi}(t)\in\mathrm{St}(N,V)$ for all times. Moreover, any stationary solution $\boldsymbol{\phi}^*\in\mathrm{St}(N,V)$ of (4.3) is the critical point of the energy $\mathcal{E}$ in (2.7), since it satisfies $\operatorname{grad}\mathcal{E}(\boldsymbol{\phi}^*)=0$.

In the following subsection, we show that the iteration (4.2) is convergent if the step size τ_n is sufficiently small.

4.1. Convergence analysis

To show that the Riemannian gradient scheme (4.2) converges, we restrict ourselves to the case of a constant step size $\tau_n \equiv \tau$. First, we collect some assumptions which guarantee the convergence as established in Theorem 4.3 below.

(A1) (Polyak-Łojasiewicz gradient inequality) For the ground state $\boldsymbol{\phi}^* \in \mathrm{St}(N,V)$, there exist $C_*, C_{\mathrm{PL}} > 0$ such that for all $\boldsymbol{\phi} \in \mathrm{St}(N,V)$ with $\|\boldsymbol{\phi} - \boldsymbol{\phi}^*\|_{a_0} \leq C_*$, it holds

$$\left|\mathcal{E}(\boldsymbol{\phi}) - \mathcal{E}(\boldsymbol{\phi}^*)\right| \leq C_{\mathrm{PL}} \left\|\operatorname{grad} \mathcal{E}(\boldsymbol{\phi})\right\|_{a_{\boldsymbol{\phi}}}^2.$$

(A2) (Descent inequality) We say that a given sequence $\left\{\boldsymbol{\phi}^{(n)}\right\} \subset \mathrm{St}(N,V)$ satisfies the descent inequality, if there exist $C_{\mathrm{D}} > 0$ and $n_{\mathrm{D}} \in \mathbb{N}$ such that for all $n \geq n_{\mathrm{D}}$,

$$\mathcal{E}\left(\boldsymbol{\phi}^{(n)}\right) - \mathcal{E}\left(\boldsymbol{\phi}^{(n+1)}\right) \geq C_{\mathrm{D}} \left\|\operatorname{grad} \mathcal{E}\left(\boldsymbol{\phi}^{(n)}\right)\right\|_{a_{\boldsymbol{\phi}^{(n)}}} \left\|\boldsymbol{\phi}^{(n+1)} - \boldsymbol{\phi}^{(n)}\right\|_{a_0}. \tag{4.4}$$

(A3) (Step size condition) For a given sequence $\left\{\boldsymbol{\phi}^{(n)}\right\} \subset \mathrm{St}(N,V)$, we say that it satisfies the step size condition, if there exist $C_{\mathrm{S}} > 0$ and $n_{\mathrm{S}} \in \mathbb{N}$ such that for all $n \geq n_{\mathrm{S}}$,

$$\left\|\boldsymbol{\phi}^{(n+1)} - \boldsymbol{\phi}^{(n)}\right\|_{a_0} \geq C_{\mathrm{S}} \left\|\operatorname{grad} \mathcal{E}\left(\boldsymbol{\phi}^{(n)}\right)\right\|_{a_{\boldsymbol{\phi}^{(n)}}}. \tag{4.5}$$

Under these assumptions, the convergence result and the convergence rate can be established by the following theorem adapted from [38]. Its proof is a straight-forward modification of Theorem 2.1 from [38] and therefore omitted here.

Theorem 4.3. *Let $\left\{\boldsymbol{\phi}^{(n)}\right\} \subset \mathrm{St}(N,V)$ be a sequence generated by the descent gradient method (4.2), which satisfies the descent condition (**A2**). If there exists an accumulation point $\boldsymbol{\phi}^* \in \mathrm{St}(N,V)$ of the sequence that satisfies the Polyak-Łojasiewicz gradient condition (**A1**), then $\boldsymbol{\phi}^*$ is the unique limit point of $\left\{\boldsymbol{\phi}^{(n)}\right\}$ with respect to $\|\cdot\|_{a_0}$. Further, if the sequence $\left\{\boldsymbol{\phi}^{(n)}\right\}$ fulfills the step size condition (**A3**), then there exist constants $c, C > 0$ such that the convergence rate can be estimated as*

$$\|\boldsymbol{\phi}^{(n)} - \boldsymbol{\phi}^*\|_{a_0} \leq C\, e^{-cn}$$

and it holds $\lim_{n\to\infty} \operatorname{grad} \mathcal{E}\left(\boldsymbol{\phi}^{(n)}\right) = 0$.

It remains to discuss the validity of the three conditions (**A1**)–(**A3**) in the considered setting. Condition (**A1**) is an assumption on the energy and depends on the particular application. The special case of the Gross–Pitaevskii equation is discussed in detail in [38]. The other two conditions can be verified under moderate constraints on the step size and suitable regularity assumptions on the energy.

Lemma 4.4 (Sufficient condition for (**A2**))**.** *Consider a sufficiently small step size $0 < \tau \leq \tau_{\max}$. Assume that the second-order derivative of the energy is bounded in the sense that*

$$D^2\mathcal{E}(\boldsymbol{\xi})[\boldsymbol{v},\boldsymbol{w}] \leq C_0 \|\boldsymbol{v}\|_{a_0} \|\boldsymbol{w}\|_{a_0} \tag{4.6}$$

for all $\boldsymbol{\xi}$ in a small neighborhood of the ground state and all $\boldsymbol{v},\boldsymbol{w} \in V$. If the iterates $\boldsymbol{\phi}^{(n)}$ given by (4.2) with the polar decomposition based retraction (3.13) are in this neighborhood, then there exists a constant $C_{\mathrm{D}} > 0$ such that the estimate (4.4) is satisfied.

Proof. For $\boldsymbol{\eta}^{(n)} = -\operatorname{grad}\mathcal{E}\big(\boldsymbol{\phi}^{(n)}\big) = -\boldsymbol{\phi}^{(n)} + \boldsymbol{\psi}^{(n)}$ with $\boldsymbol{\psi}^{(n)} \in \big(T_{\boldsymbol{\phi}^{(n)}}\mathrm{St}(N,V)\big)_a^{\perp}$, we obtain

$$a_{\boldsymbol{\phi}^{(n)}}\Big(\boldsymbol{\phi}^{(n)}, \boldsymbol{\eta}^{(n)}\Big) = -a_{\boldsymbol{\phi}^{(n)}}\Big(\boldsymbol{\eta}^{(n)}, \boldsymbol{\eta}^{(n)}\Big) = -\Big\|\boldsymbol{\eta}^{(n)}\Big\|^2_{a_{\boldsymbol{\phi}^{(n)}}}.$$

Further, it follows from Proposition 3.11 and

$$\boldsymbol{\phi}^{(n+1)} - \boldsymbol{\phi}^{(n)} = \mathcal{R}\Big(\boldsymbol{\phi}^{(n)}, \tau\,\boldsymbol{\eta}^{(n)}\Big) - \Big(\boldsymbol{\phi}^{(n)} + \tau\,\boldsymbol{\eta}^{(n)}\Big) + \tau\,\boldsymbol{\eta}^{(n)} \tag{4.7}$$

that

$$\begin{aligned}\Big\|\boldsymbol{\phi}^{(n+1)} - \boldsymbol{\phi}^{(n)}\Big\|_{a_0} &\le \Big\|\boldsymbol{\phi}^{(n+1)} - \boldsymbol{\phi}^{(n)}\Big\|_{a_{\boldsymbol{\phi}^{(n)}}} \\ &\le \tau\,\|\boldsymbol{\eta}^{(n)}\|_{a_{\boldsymbol{\phi}^{(n)}}} + \tau^2\,\Big\|\boldsymbol{\phi}^{(n)} + \tau\,\boldsymbol{\eta}^{(n)}\Big\|_{a_{\boldsymbol{\phi}^{(n)}}}\Big\|\boldsymbol{\eta}^{(n)}\Big\|^2_H.\end{aligned} \tag{4.8}$$

Using the expression $\boldsymbol{\eta}^{(n)} = -\boldsymbol{\phi}^{(n)} + \mathcal{A}^{-1}_{\boldsymbol{\phi}^{(n)}}\boldsymbol{\phi}^{(n)}\Big[\!\Big[\boldsymbol{\phi}^{(n)}, \mathcal{A}^{-1}_{\boldsymbol{\phi}^{(n)}}\boldsymbol{\phi}^{(n)}\Big]\!\Big]^{-1}_H$ and the coercivity and boundedness of the bilinear form $a_{\boldsymbol{\phi}^{(n)}}$, we can show that there exists a constant $C_1 > 0$ such that $\|\boldsymbol{\eta}^{(n)}\|_{a_{\boldsymbol{\phi}^{(n)}}} \le C_1\|\boldsymbol{\phi}^{(n)}\|_{a_{\boldsymbol{\phi}^{(n)}}}$. Then taking into account that the iterates $\boldsymbol{\phi}^{(n)}$ are in a small neighborhood of the ground state, we estimate

$$\Big\|\boldsymbol{\phi}^{(n)}\Big\|_{a_{\boldsymbol{\phi}^{(n)}}} \le C_2, \quad \Big\|\boldsymbol{\eta}^{(n)}\Big\|_{a_{\boldsymbol{\phi}^{(n)}}} \le C_1C_2, \quad \Big\|\boldsymbol{\phi}^{(n)} + \tau\,\boldsymbol{\eta}^{(n)}\Big\|_{a_{\boldsymbol{\phi}^{(n)}}} \le (1 + \tau_{\max}C_1)\,C_2 \tag{4.9}$$

with a constant $C_2 > 0$ independent of $\boldsymbol{\phi}^{(n)}$.

A Taylor expansion of $\mathcal{E}\big(\boldsymbol{\phi}^{(n+1)}\big)$ at $\boldsymbol{\phi}^{(n)}$ yields

$$\mathcal{E}\Big(\boldsymbol{\phi}^{(n+1)}\Big) = \mathcal{E}\Big(\boldsymbol{\phi}^{(n)}\Big) + \mathrm{D}\mathcal{E}\Big(\boldsymbol{\phi}^{(n)}\Big)\Big[\boldsymbol{\phi}^{(n+1)} - \boldsymbol{\phi}^{(n)}\Big] + \frac{1}{2}\,\mathrm{D}^2\mathcal{E}(\boldsymbol{\xi})\Big[\boldsymbol{\phi}^{(n+1)} - \boldsymbol{\phi}^{(n)}, \boldsymbol{\phi}^{(n+1)} - \boldsymbol{\phi}^{(n)}\Big]$$

for some $\boldsymbol{\xi}$ in the neighborhood of the ground state. Estimating the derivative

$$\begin{aligned}\mathrm{D}\mathcal{E}\Big(\boldsymbol{\phi}^{(n)}\Big)\Big[\boldsymbol{\phi}^{(n+1)} - \boldsymbol{\phi}^{(n)}\Big] &= a_{\boldsymbol{\phi}^{(n)}}\Big(\boldsymbol{\phi}^{(n)}, \boldsymbol{\phi}^{(n+1)} - \boldsymbol{\phi}^{(n)}\Big) \\ &\le \tau\,a_{\boldsymbol{\phi}^{(n)}}\Big(\boldsymbol{\phi}^{(n)}, \boldsymbol{\eta}^{(n)}\Big) + \Big\|\boldsymbol{\phi}^{(n)}\Big\|_{a_{\boldsymbol{\phi}^{(n)}}}\Big\|\mathcal{R}\Big(\boldsymbol{\phi}^{(n)}, \tau\,\boldsymbol{\eta}^{(n)}\Big) - \Big(\boldsymbol{\phi}^{(n)} + \tau\,\boldsymbol{\eta}^{(n)}\Big)\Big\|_{a_{\boldsymbol{\phi}^{(n)}}} \\ &\le -\tau\,\Big\|\boldsymbol{\eta}^{(n)}\Big\|^2_{a_{\boldsymbol{\phi}^{(n)}}} + C_H^2\,\tau^2\,\Big\|\boldsymbol{\phi}^{(n)}\Big\|_{a_{\boldsymbol{\phi}^{(n)}}}\Big\|\boldsymbol{\phi}^{(n)} + \tau\,\boldsymbol{\eta}^{(n)}\Big\|_{a_{\boldsymbol{\phi}^{(n)}}}\Big\|\boldsymbol{\eta}^{(n)}\Big\|^2_{a_{\boldsymbol{\phi}^{(n)}}}\end{aligned}$$

and using (4.6) together with (4.9), we conclude that

$$\begin{aligned}\mathcal{E}\Big(\boldsymbol{\phi}^{(n)}\Big) - \mathcal{E}\Big(\boldsymbol{\phi}^{(n+1)}\Big) &= -\mathrm{D}\mathcal{E}\Big(\boldsymbol{\phi}^{(n)}\Big)\Big[\boldsymbol{\phi}^{(n+1)} - \boldsymbol{\phi}^{(n)}\Big] - \frac{1}{2}\,\mathrm{D}^2\mathcal{E}(\boldsymbol{\xi})\Big[\boldsymbol{\phi}^{(n+1)} - \boldsymbol{\phi}^{(n)}, \boldsymbol{\phi}^{(n+1)} - \boldsymbol{\phi}^{(n)}\Big] \\ &\ge \tau\,\Big\|\boldsymbol{\eta}^{(n)}\Big\|^2_{a_{\boldsymbol{\phi}^{(n)}}} - C_H^2\,\tau^2\,\Big\|\boldsymbol{\phi}^{(n)}\Big\|_{a_{\boldsymbol{\phi}^{(n)}}}\Big\|\boldsymbol{\phi}^{(n)} + \tau\,\boldsymbol{\eta}^{(n)}\Big\|_{a_{\boldsymbol{\phi}^{(n)}}}\Big\|\boldsymbol{\eta}^{(n)}\Big\|^2_{a_{\boldsymbol{\phi}^{(n)}}} \\ &\quad - 2\,C_0\,\tau^2\Big\|\boldsymbol{\eta}^{(n)}\Big\|^2_{a_{\boldsymbol{\phi}^{(n)}}} - 2\,C_0\,C_H^2\,\tau^4\Big\|\boldsymbol{\phi}^{(n)} + \tau\boldsymbol{\eta}^{(n)}\Big\|^2_{a_{\boldsymbol{\phi}^{(n)}}}\Big\|\boldsymbol{\eta}^{(n)}\Big\|^4_{a_{\boldsymbol{\phi}^{(n)}}} \\ &\ge \tau\,\Big\|\boldsymbol{\eta}^{(n)}\Big\|^2_{a_{\boldsymbol{\phi}^{(n)}}}\,\big(1 - \tau_{\max}\,C_3 - \tau^3_{\max}C_4\big)\end{aligned}$$

with $C_3 = C_H^2C_2^2(1 + \tau_{\max}C_1) + 2\,C_0$ and $C_4 = 2\,C_0C_H^2C_1^2C_2^4(1 + \tau_{\max}C_1)^2$.

Finally, it follows from (4.8) and (4.9) that

$$\tau \left\| \boldsymbol{\eta}^{(n)} \right\|_{a_{\boldsymbol{\phi}^{(n)}}} \geq \frac{\left\| \boldsymbol{\phi}^{(n+1)} - \boldsymbol{\phi}^{(n)} \right\|_{a_0}}{1 + \tau_{\max} C_5},$$

with $C_5 = C_H^2 C_1 C_2^2 (1 + \tau_{\max} C_1)$. Thus, we obtain the estimate (4.4) for the sufficiently small step size $0 < \tau \leq \tau_{\max}$ and a constant $C_{\mathrm{D}} > 0$ depending on $\tau_{\max}$ and the other constants only. □

Lemma 4.5 (Sufficient condition for (**A3**))**.** *Consider a sufficiently small step size $0 < \tau_{\min} \leq \tau \leq \tau_{\max}$. If the iterates $\boldsymbol{\phi}^{(n)}$ given by (4.2) with the polar decomposition based retraction (3.13) are in the neighborhood of the ground state, then there exists a constant $C_{\mathrm{S}} > 0$ such that the estimate (4.5) is satisfied.*

Proof. Using (4.7) and (4.9), we estimate

$$\begin{aligned} \tau \left\| \boldsymbol{\eta}^{(n)} \right\|_{a_{\boldsymbol{\phi}^{(n)}}} &\leq \left\| \boldsymbol{\phi}^{(n+1)} - \boldsymbol{\phi}^{(n)} \right\|_{a_{\boldsymbol{\phi}^{(n)}}} + \tau^2 \left\| \boldsymbol{\phi}^{(n)} + \tau\, \boldsymbol{\eta}^{(n)} \right\|_{a_{\boldsymbol{\phi}^{(n)}}} \left\| \boldsymbol{\eta}^{(n)} \right\|_H^2 \\ &\leq \left\| \boldsymbol{\phi}^{(n+1)} - \boldsymbol{\phi}^{(n)} \right\|_{a_{\boldsymbol{\phi}^{(n)}}} + C_5 \tau^2 \left\| \boldsymbol{\eta}^{(n)} \right\|_{a_{\boldsymbol{\phi}^{(n)}}}. \end{aligned}$$

Therefore, a step size restriction $0 < \tau_{\min} \leq \tau \leq \tau_{\max}$ with sufficiently small $\tau_{\max}$ yields

$$\begin{aligned} \left\| \boldsymbol{\phi}^{(n+1)} - \boldsymbol{\phi}^{(n)} \right\|_{a_0} &\geq c_E \left\| \boldsymbol{\phi}^{(n+1)} - \boldsymbol{\phi}^{(n)} \right\|_{a_{\boldsymbol{\phi}^{(n)}}} \geq c_E \left(1 - \tau C_5\right) \tau \left\| \boldsymbol{\eta}^{(n)} \right\|_{a_{\boldsymbol{\phi}^{(n)}}} \\ &\geq c_E \left(1 - \tau_{\max} C_5\right) \tau_{\min} \left\| \boldsymbol{\eta}^{(n)} \right\|_{a_{\boldsymbol{\phi}^{(n)}}} = C_{\mathrm{S}} \left\| \operatorname{grad} \mathcal{E}\left(\boldsymbol{\phi}^{(n)}\right) \right\|_{a_{\boldsymbol{\phi}^{(n)}}} \end{aligned}$$

with $C_{\mathrm{S}} = \tau_{\min} c_E (1 - \tau_{\max} C_5) > 0$. □

Remark 4.6. Note that in Lemmas 4.4 and 4.5, the polar decomposition based retraction can be replaced by the qR-based retraction defined in (3.18) or any other second-order bounded retraction.

4.2. Step size control with a non-monotone line search

In order to accelerate the convergence of the Riemannian gradient descent method (4.2), we determine the step size by employing the non-monotone line search algorithm [39] combined with the alternating Barzilai-Borwein step size strategy as proposed in [34]. The resulting Riemannian gradient descent method is presented in Algorithm 2.

The following theorem establishes that a convergent sequence generated by this algorithm yields a stationary point.

Theorem 4.7. *Let $\left\{\boldsymbol{\phi}^{(n)}\right\}$ be a sequence generated by Algorithm 2. Then every accumulation point $\boldsymbol{\phi}^*$ of this sequence is a critical point of $\mathcal{E}$, i.e., we have* $\operatorname{grad} \mathcal{E}(\boldsymbol{\phi}^*) = 0$.

Proof. Since the retractions considered in Section 3.3 are globally defined, the result can be proved analogously to Theorem 3.3 of [20]. □

4.3. Inexact gradient descent schemes

In this subsection, we propose an inexact gradient descent method which significantly reduces the computational complexity of the iteration (4.2).

First, we establish a connection of our minimization method to the DCM method considered in [31]. Let $\boldsymbol{\phi}^* \in \mathrm{St}(N, V)$ be a critical point of $\mathcal{E}$, *i.e.*, $\operatorname{grad} \mathcal{E}(\boldsymbol{\phi}^*) = 0$. Then (4.1) yields

$$\mathcal{A}_{\boldsymbol{\phi}^*} \boldsymbol{\phi}^* = \boldsymbol{\phi}^* \left[\!\!\left[\boldsymbol{\phi}^*, \mathcal{A}_{\boldsymbol{\phi}^*}^{-1} \boldsymbol{\phi}^* \right]\!\!\right]_H^{-1}. \tag{4.10}$$

Algorithm 2. Riemannian gradient descent method with non-monotone line search.

Input: energy $\mathcal{E}$, retraction $\mathcal{R}$, initial guess $\boldsymbol{\phi}^{(0)} \in \mathrm{St}(N,V)$, $c_0 = \mathcal{E}\left(\boldsymbol{\phi}^{(0)}\right)$, $q_0 = 1$,
parameters $\alpha \in [0,1]$, $\beta, \delta \in (0,1)$, $0 < \gamma_{\min} < \gamma_{\max}$, $\gamma_0 > 0$
for $n = 0, 1, 2, \ldots$ **do**
Compute a search direction $\boldsymbol{\eta}^{(n)}$ as an approximation of $-\operatorname{grad}\mathcal{E}\left(\boldsymbol{\phi}^{(n)}\right)$.
if $n > 0$ **then**
Compute a trial step size

$$\gamma_n = \begin{cases} \dfrac{\left(\boldsymbol{s}^{(n)}, \boldsymbol{s}^{(n)}\right)_H}{\left|\left(\boldsymbol{s}^{(n)}, \boldsymbol{y}^{(n)}\right)_H\right|} & \text{for odd } \ n, \\ \dfrac{\left|\left(\boldsymbol{s}^{(n)}, \boldsymbol{y}^{(n)}\right)_H\right|}{\left(\boldsymbol{y}^{(n)}, \boldsymbol{y}^{(n)}\right)_H} & \text{for even } \ n, \end{cases}$$

where $\boldsymbol{s}^{(n)} = \boldsymbol{\phi}^{(n)} - \boldsymbol{\phi}^{(n-1)}$ and $\boldsymbol{y}^{(n)} = \boldsymbol{\eta}^{(n-1)} - \boldsymbol{\eta}^{(n)}$.
Set $\gamma_n = \max(\gamma_{\min}, \min(\gamma_n, \gamma_{\max}))$.
Find the smallest $k \in \mathbb{N}$ such that $\tau_n = \gamma_n \delta^k$ satisfies the non-monotone condition

$$\mathcal{E}\left(\mathcal{R}\left(\boldsymbol{\phi}^{(n)}, \tau_n \boldsymbol{\eta}^{(n)}\right)\right) \leq c_n - \beta\, \tau_n\, a_{\boldsymbol{\phi}^{(n)}}\left(\boldsymbol{\eta}^{(n)}, \boldsymbol{\eta}^{(n)}\right).$$

Set $\boldsymbol{\phi}^{(n+1)} = \mathcal{R}\left(\boldsymbol{\phi}^{(n)}, \tau_n \boldsymbol{\eta}^{(n)}\right)$.
Compute $q_{n+1} = \alpha q_n + 1$ and $c_{n+1} = \left(1 - \frac{1}{q_{n+1}}\right) c_n + \frac{1}{q_{n+1}} \mathcal{E}\left(\boldsymbol{\phi}^{(n+1)}\right)$.
Output: sequence of iterates $\left\{\boldsymbol{\phi}^{(n)}\right\}$

This equation further implies

$$[\![\boldsymbol{\phi}^*, \mathcal{A}_{\boldsymbol{\phi}^*} \boldsymbol{\phi}^*]\!]_H = \left[\!\left[\boldsymbol{\phi}^*, \mathcal{A}_{\boldsymbol{\phi}^*}^{-1} \boldsymbol{\phi}^*\right]\!\right]_H^{-1} \tag{4.11}$$

and, hence, (4.10) can be rewritten as $\mathcal{A}_{\boldsymbol{\phi}^*} \boldsymbol{\phi}^* = \boldsymbol{\phi}^* [\![\boldsymbol{\phi}^*, \mathcal{A}_{\boldsymbol{\phi}^*} \boldsymbol{\phi}^*]\!]_H$. In the DCM method considered in [31], the search direction is taken as

$$-\mathcal{B}_{\boldsymbol{\phi}}^{-1}\left(\mathcal{A}_{\boldsymbol{\phi}} \boldsymbol{\phi} - \boldsymbol{\phi}\, [\![\boldsymbol{\phi}, \mathcal{A}_{\boldsymbol{\phi}} \boldsymbol{\phi}]\!]_H\right), \tag{4.12}$$

where $\mathcal{B}_{\boldsymbol{\phi}}$ is a given preconditioner. Without $\mathcal{B}_{\boldsymbol{\phi}}$, this leads to the Riemannian gradient descent method in the Hilbert metric g_H, which usually shows slow convergence. Considering the preconditioner $\mathcal{B}_{\boldsymbol{\phi}} = \mathcal{A}_{\boldsymbol{\phi}}$ yields the search direction $-\boldsymbol{\phi} + \mathcal{A}_{\boldsymbol{\phi}}^{-1} \boldsymbol{\phi}\, [\![\boldsymbol{\phi}, \mathcal{A}_{\boldsymbol{\phi}} \boldsymbol{\phi}]\!]_H$. Due to (4.11), this search direction is asymptotically equivalent to

$$\boldsymbol{\eta} = -\operatorname{grad}\mathcal{E}(\boldsymbol{\phi}) = -\boldsymbol{\phi} + \mathcal{A}_{\boldsymbol{\phi}}^{-1} \boldsymbol{\phi} \left[\!\left[\boldsymbol{\phi}, \mathcal{A}_{\boldsymbol{\phi}}^{-1} \boldsymbol{\phi}\right]\!\right]_H^{-1}.$$

This observation shows that a suitably preconditioned DCM admits a near gradient descent structure in the novel metric g_a.

The computation of both search directions requires the solution of a system involving the operator $\mathcal{A}_{\boldsymbol{\phi}}$ in each step but different linear combinations of the outcome are used. For the DCM, it is known that an approximation of $\mathcal{A}_{\boldsymbol{\phi}}$ is sufficient for convergence in practice. In this spirit, we may also use the inexact gradient. This consideration motivates to use

$$-\operatorname{grad}\mathcal{E}(\boldsymbol{\phi}) \approx -\boldsymbol{\phi} + \mathcal{B}_{\boldsymbol{\phi}}^{-1} \boldsymbol{\phi} \left[\!\left[\boldsymbol{\phi}, \mathcal{B}_{\boldsymbol{\phi}}^{-1} \boldsymbol{\phi}\right]\!\right]_H^{-1} \tag{4.13}$$

as a search direction. Here, $\mathcal{B}_{\boldsymbol{\phi}} \approx \mathcal{A}_{\boldsymbol{\phi}}$ is a suitable preconditioner that realizes, *e.g.*, a few iterations of a preconditioned iterative solver for $\mathcal{A}_{\boldsymbol{\phi}}^{-1} \boldsymbol{\phi}$ with starting value

$$\boldsymbol{\phi}\, [\![\boldsymbol{\phi}, \mathcal{A}_{\boldsymbol{\phi}} \boldsymbol{\phi}]\!]_H^{-1} \approx \mathcal{A}_{\boldsymbol{\phi}}^{-1} \boldsymbol{\phi}.$$

The error of the proposed starting value is roughly as accurate as the current approximation of the wavefunction in the iteration. Hence, after only a few steps of the preconditioned iterative solver the residual of the linear system is substantially smaller than the current error. In a convergent iteration, sufficiently many (inner) iterations will guarantee that the resulting direction is a descent direction, *cf.* the numerical experiments of Section 5. The control of the number of iterations required to ensure a descent could be integrated into the method.

5. Examples

In this final section, we present two examples which fit in the framework of Section 2. Moreover, the efficiency of the proposed algorithm (and its preconditioned variants) are illustrated in a number of numerical experiments.

5.1. Gross–Pitaevskii eigenvalue problem

In the special case $N = 1$, we seek an eigenfunction $u \in V := H_0^1(\Omega)$ satisfying the normalization constraint $\|u\|_{L^2(\Omega)} = 1$. Hence, the minimization takes place on the unit sphere $\mathbb{S} = \{v \in V \,:\, \|u\|_{L^2(\Omega)} = 1\}$. A well-known example, which fits in this framework, is the *Gross–Pitaevskii eigenvalue problem*. In the classical form, this reads

$$-\Delta u + V_{\text{ext}}\, u + \kappa\, |u|^2 u = \lambda\, u$$

for some non-negative and space-dependent external potential $V_{\text{ext}} \geq 0$ and a constant $\kappa \geq 0$ regulating the strength of the nonlinearity. Here, the bilinear form $a_u \colon V \times V \to \mathbb{R}$ is given by

$$a_u(v, w) := \int_\Omega \nabla v \cdot \nabla w + V_{\text{ext}}\, v\, w + \kappa\, |u|^2 v\, w \,\mathrm{d}x.$$

The linear part a_0, which contains the weak Laplacian and the potential, defines an inner product on V. For the nonlinear part, we set $\gamma(\rho(u)) = \gamma(|u|^2) := \kappa\, |u|^2$, *i.e.*, a constant times the density of u. Hence, for any $u \in V$, the bilinear form a_u defines an inner product on V and Assumption 2.2 is satisfied. Due to $\Gamma(\rho) = \kappa \int_0^\rho t \,\mathrm{d}t = \frac{1}{2}\kappa\rho^2$, the corresponding energy has the form

$$\mathcal{E}(u) = \frac{1}{2}\, a_0(u, u) + \frac{1}{2} \int_\Omega \Gamma(\rho(u)) \,\mathrm{d}x = \frac{1}{2} \int_\Omega \|\nabla u\|^2 + V_{\text{ext}}\, |u|^2 + \frac{\kappa}{2}\, |u|^4 \,\mathrm{d}x.$$

The assumed property that $\mathcal{E}$ does not change if the argument is multiplied by an orthogonal matrix translates in the case $N = 1$ to $\mathcal{E}(\pm u) = \mathcal{E}(u)$, which is clearly satisfied. As before, we are interested in the ground state, *i.e.*, the state of minimal energy. For the Gross–Pitaevskii eigenvalue problem, the ground state coincides with the eigenfunction that corresponds to the smallest eigenvalue.

Following the procedure presented in Section 2, we have $[\![u, v]\!]_H = (u, v)_H = (u, v)_{L^2(\Omega)}$ and $T_u\,\mathbb{S} = \{v \in V \,:\, (u, v)_H = 0\}$. Hence, the normal space is one-dimensional, and (3.4) reduces to find $\psi \in V$ such that

$$a_u(\psi, v) = 0 \quad \text{for all } v \in T_u\,\mathbb{S}, \qquad (\psi, u)_H = 1.$$

Written as a saddle point problem, we seek $(\psi, \mu) \in V \times \mathbb{R}$ such that

$$a_u(\psi, v) = \mu\, (u, v)_H \qquad \text{for all } v \in V, \tag{5.1a}$$

$$(\psi, u)_H = 1. \tag{5.1b}$$

The resulting projection applied to u reads $P_u(u) = u - \psi$. For $N = 1$, the polar decomposition based retraction from Section 3.3.1 as well as the qR-based retraction from Section 3.3.2 simply equal a L^2-normalization. This then leads to the following iteration scheme: Given $u^{(n)} \in \mathbb{S}$, compute $\psi^{(n)} = \psi(u^{(n)})$ by solving (5.1) with $u = u^{(n)}$ and set

$$\tilde{u}^{(n+1)} := (1 - \tau_n) u^{(n)} + \tau_n \psi^{(n)}, \qquad u^{(n+1)} := \frac{\tilde{u}^{(n+1)}}{\big\|\tilde{u}^{(n+1)}\big\|_{L^2(\Omega)}}.$$

Note that this is exactly the damped GFa_z method introduced in [18], which is labeled A-method in [5]. Moreover, in the special case $\tau_n \equiv 1$, this iteration is the straight-forward generalization of the *inverse power method* to the nonlinear setting. We refer to the aforementioned original papers as well as to [4,6] for numerical experiments that demonstrate the competitiveness of the method with established schemes and its ability to capture relevant physical phenomena such as the exponential localization of eigenstates. The guaranteed energy decay of the method has also been exploited explicitly in [13].

5.2. Kohn–Sham model

A second example, which is covered by this paper, is the Kohn–Sham model [23] and, in particular, the model based on the *density functional theory* [19]. This theory allows a reduction of the degrees of freedom, leading to a model which balances accuracy and computational cost, see also [11,12,36] for a more detailed introduction.

5.2.1. Validation of the model

As an energy functional, we consider (with $\Omega = \mathbb{R}^3$)

$$\begin{aligned}\mathcal{E}(\boldsymbol{\phi}) = \frac{1}{2}\sum_{j=1}^{N}\int_\Omega \|\nabla\phi_j(r)\|^2\,\mathrm{d}r + \int_\Omega V_{\mathrm{ion}}(r)\,\rho(\boldsymbol{\phi}(r))\,\mathrm{d}r \\ + \frac{1}{2}\int_\Omega\int_\Omega \frac{\rho(\boldsymbol{\phi}(r))\,\rho(\boldsymbol{\phi}(r'))}{\|r-r'\|}\,\mathrm{d}r\,\mathrm{d}r' + \int_\Omega \epsilon_{\mathrm{xc}}(\rho(\boldsymbol{\phi}(r)))\,\rho(\boldsymbol{\phi}(r))\,\mathrm{d}r\end{aligned} \tag{5.2}$$

with the ionic potential V_{ion}, the exchange-correlation ϵ_{xc}, and the associated electronic charge density $\rho(\boldsymbol{\phi}(r)) = \boldsymbol{\phi}(r)\cdot\boldsymbol{\phi}(r) = \sum_{j=1}^{N}|\phi_j(r)|^2$. Based on semi-empirically knowledge of the model, the particular exchange-correlation is described in [36]. For more details, on this and the corresponding *local density approximation*, we refer to [27,28]. Following the physical setup, the ionic potential typically reads

$$V_{\mathrm{ion}}(r) = \sum_{j=1}^{N_{\mathrm{nuc}}}\frac{z_j}{\|r-r_j\|}$$

with the number of nuclei N_{nuc}, the charge of the jth nuclei z_j, and its position r_j, which are assumed to be fixed. The obvious problem of the included singularities can be circumvented by considering core electrons (which are very close to a nucleus) as part of the corresponding core. For more details on this so-called *pseudopotential approximation*, we refer once more to [36] and the references therein. As a consequence, we may assume in the following that V_{ion} in (5.2) is a bounded potential.

We are interested in the Kohn–Sham ground state, which means that we aim to minimize the energy $\mathcal{E}$ over $V = \tilde{V}^N$ with $\tilde{V} = H^1_{\mathrm{per}}(\Omega)$, *i.e.*, the Sobolev space $H^1(\Omega)$ with periodic boundary conditions, subject to the constraint $[\![\boldsymbol{\phi},\boldsymbol{\phi}]\!]_H = \boldsymbol{I}_N$. Hence, the minimization takes place on the Stiefel manifold $\mathrm{St}(N,V)$. Following (2.3), the corresponding bilinear form reads

$$a_{\boldsymbol{\phi}}(\boldsymbol{v},\boldsymbol{w}) = \int_\Omega \mathrm{tr}\Big((\nabla\boldsymbol{v})^T\nabla\boldsymbol{w}\Big)\,\mathrm{d}r + 2\int_\Omega V_{\mathrm{ion}}\,\boldsymbol{v}\cdot\boldsymbol{w}\,\mathrm{d}r + \int_\Omega \gamma(\rho(\boldsymbol{\phi}))\,\boldsymbol{v}\cdot\boldsymbol{w}\,\mathrm{d}r$$

with the (non-local) nonlinearity

$$\gamma(\rho) = 2\int_\Omega \frac{\rho(\boldsymbol{\phi}(r'))}{\|r-r'\|}\,\mathrm{d}r' + 2\,\frac{\mathrm{d}}{\mathrm{d}\rho}(\rho\,\epsilon_{\mathrm{xc}}(\rho)).$$

Lemma 5.1. *Consider a fixed $\boldsymbol{\phi}\in V$ and assume that V_{ion} and $\gamma(\rho(\boldsymbol{\phi}))$ are bounded. Then the corresponding bilinear form*

$$\tilde{a}_{\boldsymbol{\phi}}(v_j,w_j) = \int_\Omega (\nabla v_j)^T\nabla w_j\,\mathrm{d}r + 2\int_\Omega V_{\mathrm{ion}}\,v_j\,w_j\,\mathrm{d}r + \int_\Omega \gamma(\rho(\boldsymbol{\phi}))\,v_j\,w_j\,\mathrm{d}r$$

satisfies a Gårding inequality. Hence, there exists $\sigma\in\mathbb{R}$ such that $\tilde{a}_{\boldsymbol{\phi}} + \sigma\,(\cdot,\cdot)_{L^2(\Omega)}$ is a symmetric, bounded, and coercive bilinear form.

Proof. Let c_V and c_γ denote the (possibly negative) lower bounds of $2V_{\text{ion}}$ and $\gamma(\rho(\boldsymbol{\phi}))$, respectively. Then, the definition of $\tilde{a}_{\boldsymbol{\phi}}$ gives

$$\tilde{a}_{\boldsymbol{\phi}}(v,v) \geq \int_\Omega \|\nabla v\|^2 \,\mathrm{d}r + (c_V + c_\gamma)\,(v,v)_{L^2(\Omega)} = \|v\|_V^2 - (1 - c_V - c_\gamma)\,\|v\|_{L^2(\Omega)}^2$$

for all $v \in \tilde{V}$. The coercivity of $\tilde{a}_{\boldsymbol{\phi}} + \sigma\,(\cdot,\cdot)_{L^2(\Omega)}$ then follows for any $\sigma \geq 1 - c_V - c_\gamma$. Symmetry and boundedness are directly given. □

Since we cannot ensure that $\tilde{a}_{\boldsymbol{\phi}}$ is coercive, we need to adapt the original nonlinear eigenvector problem (2.9) by a *shift*: seek $\boldsymbol{\phi} \in \mathrm{St}(N,V)$ and $\lambda_1, \ldots, \lambda_N \in \mathbb{R}$ such that

$$\tilde{a}_{\boldsymbol{\phi}}(\phi_j, v_j) + \sigma\,(\phi_j, v_j)_{L^2(\Omega)} = (\lambda_j + \sigma)\,(\phi_j, v_j)_{L^2(\Omega)} \qquad \text{for all } (v_1, \ldots, v_N) \in V$$

with the shift σ from Lemma 5.1. This gives a coupled system of nonlinear eigenvector problems, which satisfies Assumption 2.2 and, therefore, the theory of this paper is applicable.

5.2.2. Numerical experiments

We now illustrate the convergence behaviour of the new energy-adaptive Riemannian gradient descent scheme (RGD) and its variants and show that they are competitive with the established SCF iteration and the preconditioned DCM method. The numerical experiments are performed on an Intel(R) Core(TM) i7-8565U CPU@1.80 GHz using MATLAB (version R2021b). The implementation is based on the MATLAB toolbox KSSOLV, *cf.* [36]. The usage of this toolbox allows us to focus on the new eigenvalue iterations and their comparison to already existing methods. Note that the toolbox works with an additional factor of two in the electronic charge density. This, however, does not affect the convergence behaviour. We initially select an exemplary molecule system implemented in KSSOLV, namely CO_2 ($N = 8$). In KSSOLV, a spatial discretization using a planewave discretization of functions in $V := [H^1(\mathbb{R}^3)]^N$ is considered. As in [36], we use a $32 \times 32 \times 32$ sampling grid for the wavefunctions in the CO_2 model.

We shall first illustrate the convergence behaviour of the RGD. For the CO_2 molecule, we compare the following variants:

- RGD from (4.2) for several choices of a constant time step size $\tau_n = \tau$ with $\tau \in \{0.05, 0.1, 0.15, 0.2\}$,
- RGD with the non-monotone line search as presented in Algorithm 2 with the descent direction $\boldsymbol{\eta}^{(n)} = -\operatorname{grad}\mathcal{E}\big(\boldsymbol{\phi}^{(n)}\big)$ and parameters $\alpha = 0.95$, $\beta = 10^{-4}$, $\gamma_{\min} = 10^{-4}$, $\gamma_{\max} = 1.0$, $\gamma_0 = 10^{-2}$, and $\delta = 0.5$.

For both variants, the polar decomposition based retraction (3.13) is used. As a stopping criterion, we consider the H-norm of the residual to fall below the tolerance $\mathrm{tol} = 10^{-6}$. The linear systems are solved up to the higher accuracy of 10^{-8}.

Figure 1 (left) shows the evolution of the residuals in the iteration. In accordance with the theoretical predictions, we observe convergence for sufficiently small constant step sizes. A look into the corresponding errors in the energy (with respect to a reference minimal energy computed to higher accuracy) depicted in Figure 1 (right) shows that for $\tau = 0.2$, after an initial decay, the method approaches some other critical point on a higher energy level. Furthermore, for smaller choices of τ, the linear convergence to the ground state is observed. There is probably an optimal choice of the step size around $\tau = 0.15$ that minimizes the linear rate of convergence. However, the non-monotone line search converges much faster and appears to be much more efficient for this example and many others that we have tried.

While the line search optimizes the iteration count, the cost per iteration step is largely reduced by the inexact solution of linear system for the gradient computation in each step. We will refer to the corresponding scheme as:

- inexact RGD with the non-monotone line search as presented in Algorithm 2 with $\boldsymbol{\eta}^{(n)}$ being the preconditioned MINRES approximation given in (4.13). We use the MINRES implementation of MATLAB using the KSSOLV built-in Teter preconditioner [32,36].

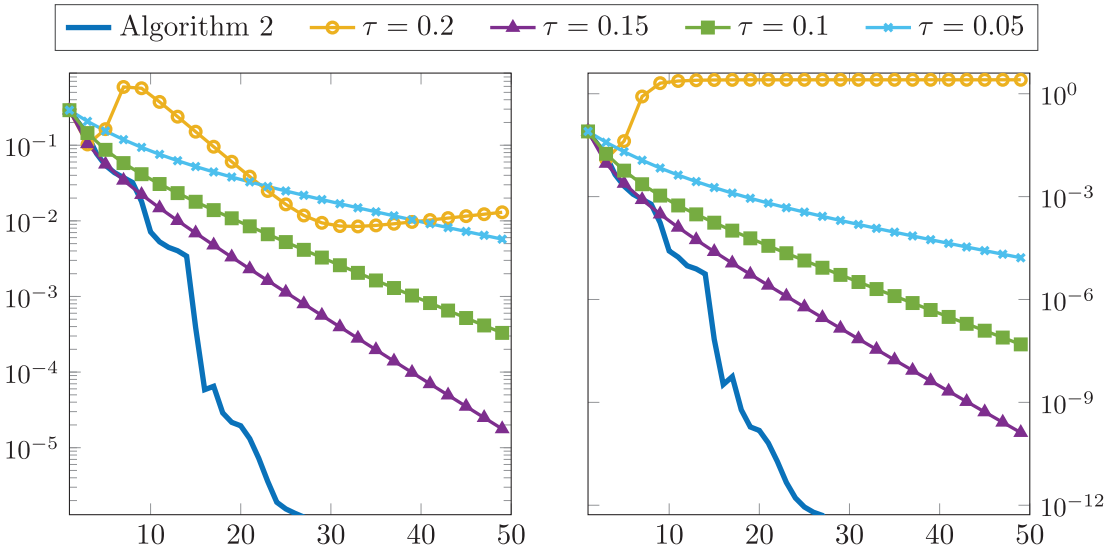

FIGURE 1. Convergence history of the residual (*left*) and energy (*right*) for the CO_2 model for different (fixed) step sizes and the non-monotone line search from Algorithm 2.

TABLE 1. CPU time (in seconds) and number of needed iteration steps to achieve an approximation of the ground state of the CO_2 molecule with tolerance 10^{-6} in the residual.

	SCF	RGD	inexact RGD	prec. DCM
CPU time	12.3	36.7	11.6	16.6
# iterations	8	28	37	45

We also compare the performance of the exact and inexact RGD with the established schemes

- SCF: self-consistent field iteration as readily available in KSSOLV using LOPCG to solve the linear eigenvalue problem in each step up to tolerance 10^{-8},
- DCM: direct constrained minimization as defined in (4.12) with non-monotone line search. The preconditioner is given by 3 steps of the preconditioned MINRES iteration as in the inexact RGD.

All schemes use the same initial guess to the wavefunction and the qR-based retraction defined in (3.18). For the RDG variants and DCM, we use the non-monotone line search with the prescribed parameters given above.

Table 1 shows the CPU times and (outer) iteration counts of the four methods. While solving the linear systems too accurately seems to be suboptimal in terms of computational complexity, the numbers clearly indicate that the inexact RGD substantially accelerates the simulation and is very competitive with SCF. The closely related preconditioned DCM variant performs equally well asymptotically but, according to our experience, is a bit slower in the initial phase when the residuals are still large.

According to our experience, the competitiveness of inexact RGD is representative. An experiment for the more challenging molecule petacene, also implemented in KSSOLV, supports this assessment; see Table 2. Due to the large number of electrons in pentacene ($N = 102$), a sampling grid of size $64 \times 32 \times 48$ is used for the spatial planewave discretization.

TABLE 2. CPU time (in seconds) and number of needed iteration steps to achieve an approximation of the ground state of the *pentacene* molecule with tolerance 10^{-6} in the residual.

	SCF	inexact RGD	prec. DCM
CPU time	3211	2204	2655
# iterations	14	43	51

6. Conclusion

In this paper, we have generalized the energy-adaptive gradient descent scheme from [18] to nonlinear eigenvector problems formulated on the infinite-dimensional Stiefel manifold. We have shown convergence of the method and a guaranteed energy decay of the iterates if the step size is sufficiently small. Moreover, we have introduced a non-monotone step size control and discussed the inexact variants, which accelerate the proposed method significantly. In total, this gives a novel energy-adaptive descent scheme, which is competitive with existing schemes such as SCF and DCM.

Appendix A. Proofs of second-order bounds for the retractions

A.1. Proof of Proposition 3.11

For

$$\boldsymbol{w}(t) := \mathcal{R}(\boldsymbol{\phi}, t\boldsymbol{\eta}) - (\boldsymbol{\phi} + t\boldsymbol{\eta}) = (\boldsymbol{\phi} + t\boldsymbol{\eta})\Big(\big(\boldsymbol{I}_N + t^2[\![\boldsymbol{\eta}, \boldsymbol{\eta}]\!]_H\big)^{-1/2} - \boldsymbol{I}_N\Big)$$

we have

$$\|\boldsymbol{w}(t)\|_{a_{\boldsymbol{\phi}}} \leq \|\boldsymbol{\phi} + t\boldsymbol{\eta}\|_{a_{\boldsymbol{\phi}}} \left\|\big(\boldsymbol{I}_N + t^2[\![\boldsymbol{\eta}, \boldsymbol{\eta}]\!]_H\big)^{-1/2} - \boldsymbol{I}_N\right\|_2,$$

where $\|\cdot\|_2$ denotes the spectral matrix norm. Let $\mu_1 \geq \ldots \geq \mu_N \geq 0$ be the eigenvalues of the symmetric, positive semidefinite matrix $[\![\boldsymbol{\eta}, \boldsymbol{\eta}]\!]_H$. Then,

$$\left\|\big(\boldsymbol{I}_N + t^2[\![\boldsymbol{\eta}, \boldsymbol{\eta}]\!]_H\big)^{-1/2} - \boldsymbol{I}_N\right\|_2 = \max_{1\leq j\leq N}\left(1 - \frac{1}{\sqrt{1+t^2\mu_j}}\right) = 1 - \frac{1}{\sqrt{1+t^2\mu_1}}.$$

By the mean value theorem, there exists $\theta \in (0, t)$ such that

$$1 - \frac{1}{\sqrt{1+t^2\mu_1}} = \frac{\theta\, t\, \mu_1}{\sqrt{(1+\theta^2\mu_1)^3}}.$$

This implies

$$\left\|\big(\boldsymbol{I}_N + t^2[\![\boldsymbol{\eta}, \boldsymbol{\eta}]\!]_H\big)^{-1/2} - \boldsymbol{I}_N\right\|_2 \leq t^2\mu_1 = t^2\,\|[\![\boldsymbol{\eta}, \boldsymbol{\eta}]\!]_H\|_2 \leq t^2\,\|\boldsymbol{\eta}\|_H^2.$$

Thus, the assertion holds true. □

A.2. Proof of Proposition 3.13

For a curve $\boldsymbol{\varphi}(t) = \boldsymbol{\phi} + t\boldsymbol{\eta}$, consider the qR decomposition $\boldsymbol{\varphi}(t) = \boldsymbol{q}(t)R(t)$. Then we have

$$\begin{aligned}\|\mathcal{R}(\boldsymbol{\phi}, t\boldsymbol{\eta}) - (\boldsymbol{\phi} + t\boldsymbol{\eta})\|_{a_{\boldsymbol{\phi}}} &= \|\boldsymbol{q}(t) - \boldsymbol{q}(t)R(t)\|_{a_{\boldsymbol{\phi}}} \leq \|\boldsymbol{q}(t)\|_{a_{\boldsymbol{\phi}}}\|R(0) - R(t)\|_2 \\ &\leq \|\boldsymbol{\phi} + t\boldsymbol{\eta}\|_{a_{\boldsymbol{\phi}}}\big\|R^{-1}(t)\big\|_2 \int_0^t \left\|\dot{R}(s)\right\|_F \mathrm{d}s. \end{aligned} \tag{A.1}$$

Differentiating the relation

$$\boldsymbol{I}_N + t^2 [\![\boldsymbol{\eta}, \boldsymbol{\eta}]\!]_H = [\![\boldsymbol{\phi} + t\boldsymbol{\eta}, \boldsymbol{\phi} + t\boldsymbol{\eta}]\!]_H = R^T(t)R(t), \tag{A.2}$$

we obtain

$$2t\, [\![\boldsymbol{\eta}, \boldsymbol{\eta}]\!]_H = \dot{R}^T(t)R(t) + R^T(t)\dot{R}(t).$$

Multiplying this equation by $R^{-T}(t)$ and $R^{-1}(t)$ from the left and right, respectively, yields

$$\big(\dot{R}(t)R^{-1}(t)\big)^T + \dot{R}(t)R^{-1}(t) = 2t\, R^{-T}(t)[\![\boldsymbol{\eta}, \boldsymbol{\eta}]\!]_H R^{-1}(t).$$

Since $\dot{R}(t)R^{-1}(t)$ is upper triangular, we obtain

$$\dot{R}(t) = 2t\, \mathrm{up}\big(R^{-T}(t)[\![\boldsymbol{\eta}, \boldsymbol{\eta}]\!]_H R^{-1}(t)\big)R(t),$$

where

$$(\mathrm{up}(M))_{ij} = \begin{cases} M_{ij}, & \text{if } 1 \le i < j \le N, \\ \frac{1}{2}M_{ij}, & \text{if } 1 \le i = j \le N, \\ 0, & \text{otherwise} \end{cases}$$

for any symmetric matrix $M = [M_{ij}] \in \mathbb{R}^{N\times N}$. As before, let $\mu_1 \ge \ldots \ge \mu_N \ge 0$ denote the eigenvalues of $[\![\boldsymbol{\eta}, \boldsymbol{\eta}]\!]_H$. Using $2\,\|\mathrm{up}(M)\|_F^2 \le \|M\|_F^2$ and (A.2), we have

$$\begin{aligned} 2\big\|\mathrm{up}\big(R^{-T}(t)[\![\boldsymbol{\eta}, \boldsymbol{\eta}]\!]_H R^{-1}(t)\big)\big\|_F^2 &\le \big\|R^{-T}(t)[\![\boldsymbol{\eta}, \boldsymbol{\eta}]\!]_H R^{-1}(t)\big\|_F^2 \\ &= \mathrm{tr}\big([\![\boldsymbol{\eta}, \boldsymbol{\eta}]\!]_H R^{-1}(t)R^{-T}(t)[\![\boldsymbol{\eta}, \boldsymbol{\eta}]\!]_H R^{-1}(t)R^{-T}(t)\big) \\ &= \Big\|[\![\boldsymbol{\eta}, \boldsymbol{\eta}]\!]_H \big(\boldsymbol{I}_N + t^2[\![\boldsymbol{\eta}, \boldsymbol{\eta}]\!]_H\big)^{-1}\Big\|_F^2 \\ &= \sum_{i=1}^N \left(\frac{\mu_i}{1+t^2\mu_i}\right)^2 \le \sum_{i=1}^N \mu_i^2 = \|[\![\boldsymbol{\eta}, \boldsymbol{\eta}]\!]_H\|_F^2 \le \|\boldsymbol{\eta}\|_H^4 \end{aligned}$$

and, hence,

$$\|\dot{R}(t)\|_F \le 2\,t\|\mathrm{up}(R^{-T}(t)[\![\boldsymbol{\eta}, \boldsymbol{\eta}]\!]_H R^{-1}(t))\|_F \|R(t)\|_2 \le \sqrt{2}\,t\,\|\boldsymbol{\eta}\|_H^2 \|R(t)\|_2. \tag{A.3}$$

Furthermore, (A.2) implies that

$$|R(t)\|_2 = \big\|\boldsymbol{I}_N + t^2[\![\boldsymbol{\eta}, \boldsymbol{\eta}]\!]_H\big\|_2^{1/2} \le \left(1 + t^2\|\boldsymbol{\eta}\|_H^2\right)^{1/2}, \tag{A.4}$$

$$\big\|R^{-1}(t)\big\|_2 = \Big\|\big(\boldsymbol{I}_N + t^2[\![\boldsymbol{\eta}, \boldsymbol{\eta}]\!]_H\big)^{-1}\Big\|_2^{1/2} < 1. \tag{A.5}$$

Thus, the claimed estimate follows from (A.1), (A.3), (A.4), and (A.5). □

Acknowledgements. The authors would like to thank Patrick Henning for his inspiring work on the energy-adaptive Riemannian gradient descent method for the Gross–Pitaevskii problem and Benjamin Stamm for raising the question of its applicability to the Kohn–Sham model. The work of Daniel Peterseim is part of a project that has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (Grant agreement No. 865751 – RandomMultiScales).

References

[1] P.-A. Absil and J. Malick, Projection-like retractions on matrix manifolds. *SIAM J. Optim.* **22** (2012) 135–158.

[2] P.-A. Absil, R. Mahony and R. Sepulchre, Optimization Algorithms on Matrix Manifolds. Princeton University Press, Princeton, NJ (2008).

[3] F. Alouges and C. Audouze, Preconditioned gradient flows for nonlinear eigenvalue problems and application to the Hartree–Fock functional. *Numer. Meth. Part. D. E.* **25** (2009) 380–400.

[4] R. Altmann and D. Peterseim, Localized computation of eigenstates of random Schrödinger operators. *SIAM J. Sci. Comput.* **41** (2019) B1211–B1227.

[5] R. Altmann, P. Henning and D. Peterseim, The *J*-method for the Gross–Pitaevskii eigenvalue problem. *Numer. Math.* **148** (2021) 575–610.

[6] R. Altmann, P. Henning and D. Peterseim, Localization and delocalization of ground states of Bose–Einstein condensates under disorder. *SIAM J. Appl. Math.* **82** (2022) 330–358.

[7] W. Bao and Q. Du, Computing the ground state solution of Bose–fEinstein condensates by a normalized gradient flow. *SIAM J. Sci. Comput.* **25** (2004) 1674–1697.

[8] D. Braess, Finite Elements – Theory, Fast Solvers, and Applications in Solid Mechanics, 3rd edition. Cambridge University Press, New York (2007).

[9] E. Cancès, Self-consistent field algorithms for Kohn–Sham models with fractional occupation numbers. *J. Chem. Phys.* **114** (2001) 10616–10622.

[10] E. Cancès and C. Le Bris, On the convergence of SCF algorithms for the Hartree–Fock equations. *ESAIM: M2AN* **34** (2000) 749–774.

[11] E. Cancès, R. Chakir and Y. Maday, Numerical analysis of the planewave discretization of some orbital-free and Kohn–Sham models. *ESAIM: M2AN* **46** (2012) 341–388.

[12] E. Cancès, G. Dusson, Y. Maday, B. Stamm and M. Vohralík, A perturbation-method-based post-processing for the planewave discretization of Kohn–Sham models. *J. Comput. Phys.* **307** (2016) 446–459.

[13] E. Cancès, G. Kemlin and A. Levitt, Convergence analysis of direct minimization and self-consistent iterations. *SIAM J. Matrix Anal. Appl.* **42** (2021) 243–274.

[14] A. Edelman, T.A. Arias and S.T. Smith, The geometry of algorithms with orthogonality constraints. *SIAM J. Matrix Anal. Appl.* **20** (1998) 303–353.

[15] G.H. Golub and C.F. Van Loan, Matrix Computations, 4th edition. The Johns Hopkins University Press, Baltimore, London (2013).

[16] P. Harms and A. Mennucci, Geodesics in infinite dimensional Stiefel and Grassmann manifolds. *C. R. Math.* **350** (2012) 773–776.

[17] P. Heid, B. Stamm and T.P. Wihler, Gradient flow finite element discretizations with energy-based adaptivity for the Gross–Pitaevskii equation. *J. Comput. Phys.* **436** (2021) 110165.

[18] P. Henning and D. Peterseim, Sobolev gradient flow for the Gross–Pitaevskii eigenvalue problem: global convergence and computational efficiency. *SIAM J. Numer. Anal.* **58** (2020) 1744–1772.

[19] P. Hohenberg and W. Kohn, Inhomogeneous electron gas. *Phys. Rev.* **136** (1964) B864–B871.

[20] J. Hu, X. Liu, Z.-W. Wen and Y.-X. Yuan, A brief introduction to manifold optimization. *J. Oper. Res. Soc. China* **8** (2020) 199–248.

[21] E. Jarlebring and P. Upadhyaya, Implicit algorithms for eigenvector nonlinearities. *Numer. Algorithms* **90** (2022) 301–321.

[22] T. Kaneko, S. Fiori and T. Tanaka, Empirical arithmetic averaging over the compact Stiefel manifold. *IEEE Trans. Signal Proces.* **61** (2013) 883–894.

[23] W. Kohn and L.J. Sham, Self-consistent equations including exchange and correlation effects. *Phys. Rev.* **140** (1965) A1133–A1138.

[24] P. Lancaster and M. Tismenetsky, The Theory of Matrices, 2nd edition. Academic Press, Orlando, FL (1985).

[25] C. Le Bris, Computational chemistry from the perspective of numerical analysis. *Acta Numer.* **14** (2005) 363–444.

[26] E.H. Lieb, R. Seiringer and J. Yngvason, A rigorous derivation of the Gross–Pitaevskii energy functional for a two-dimensional Bose gas. *Comm. Math. Phys.* **224** (2001) 17–31.

[27] J.M. MacLaren, D.P. Clougherty, M.E. McHenry and M.M. Donovan, Parameterised local spin density exchange-correlation energies and potentials for electronic structure calculations I. Zero temperature formalism. *Comput. Phys. Commun.* **66** (1991) 383–391.

[28] J.P. Perdew and A. Zunger, Self-interaction correction to density-functional approximations for many-electron systems. *Phys. Rev. B* **23** (1981) 5048–5079.

[29] L.P. Pitaevskii and S. Stringari, Bose–Einstein Condensation. Oxford University Press, Oxford (2003).

[30] H. Sato and K. Aihara, Cholesky QR-based retraction on the generalized Stiefel manifold. *Comput. Optim. Appl.* **72** (2019) 293–308.

[31] R. Schneider, T. Rohwedder, A. Neelov and J. Blauert, Direct minimization for calculating invariant subspaces in density functional computations of the electronic structure. *J. Comput. Math.* **27** (2009) 360–387.

[32] M.P. Teter, M.C. Payne and D.C. Allan, Solution of Schrödinger's equation for large systems. *Phys. Rev. B* **40** (1989) 12255–12263.

[33] A. Uschmajew, Well-posedness of convex maximization problems on Stiefel manifolds and orthogonal tensor product approximations. *Numer. Math.* **115** (2010) 309–331.
[34] Z. Wen and W. Yin, A feasible method for optimization with orthogonality constraints. *Math. Program.* **142** (2013) 397–434.
[35] C. Yang, J.C. Meza and L.-W. Wang, A constrained optimization algorithm for total energy minimization in electronic structure calculation. *J. Comput. Phys.* **217** (2006) 709–721.
[36] C. Yang, J.C. Meza, B. Lee and L.-W. Wang, KSSOLV – a MATLAB toolbox for solving the Kohn–Sham equations. *ACM Trans. Math. Softw.* **36** (2009) 1–35.
[37] E. Zeidler, Nonlinear Functional Analysis and its Applications IIa: Linear Monotone Operators. Springer-Verlag, New York (1990).
[38] Z. Zhang. Exponential convergence of Sobolev gradient descent for a class of nonlinear eigenproblems. *Commun. Math. Sci.* **20** (2022) 377–403.
[39] H. Zhang and W.W. Hager, A nonmonotone line search technique and its application to unconstrained optimization. *SIAM J. Optim.* **14** (2004) 1043–1056.