SIAM J. NUMER. ANAL.
Vol. 63, No. 6, pp. 2343–2370

GENERALIZED GENTLEST ASCENT DYNAMICS METHODS FOR HIGH-INDEX SADDLE POINTS*

MOODY T. CHU† AND MATTHEW M. LIN‡

Abstract. A geometric perspective on the gentlest ascent dynamics is presented, revealing that the dynamics is utilizing the Householder reflector—constructed via the continuous power method—to adapt the negative gradient and identify index-1 saddle points. While the adaptation appears intuitive, it is governed by a precise criterion. Building on this geometric insight, three generalized dynamical systems are introduced for locating high-index saddle points, each centered on estimating directions for constructing generalized reflectors. The first approach employs the Oja flow to evolve eigenspaces, encompassing the continuous power method as a special case. The second approach formulates a matrix Riccati differential equation for the projector operator on the Grassmann manifold, which is shown to be equivalent to a double bracket flow with inherent sorting properties. The third approach is a hybrid method based on conventional subspace iteration, incorporating QR factorization for normalization. The equilibrium points of all three systems are classified, and convergence analyses are provided. These dynamical systems are readily solvable by using high-precision numerical ODE integrators. Numerical experiments confirm the theoretical results.

Key words. GAD algorithm, high-index saddle point, Householder reflector, Oja flow, double bracket flow, subspace iteration

MSC codes. 37D05, 37F05, 37B35, 37C29

DOI. 10.1137/24M1710905

1. Introduction.

The notion of saddle points is ubiquitous. The most fundamental definition in calculus about a saddle point of a smooth function is that it is a stationary point, but is not a local extremum. The significance of saddle points, however, extends well beyond this basic definition. In two-player zero-sum games over continuous spaces, for instance, a saddle point represents an equilibrium: each player adopts a strategy that simultaneously minimizes their maximum possible loss and maximizes their minimum possible gain. In convex optimization with inequality constraints, the classical saddle point theorem reveals a deep connection between the saddle points of the Lagrangian and solutions to the primal problem [7, 35], underpinning many modern optimization methods. In the setting of a complex system, some rare yet significant events of transmutations can occur between the otherwise long-lived metastable states due to the reaction dynamics. Metastability, in this context, is not simply the indefinite persistence of a state, but also is characterized by a subtle, almost hidden instability that permits eventual escape. An application exemplifying this delicate balance is the reaction-diffusion equation with a double-well potential, a model that arises in quantum mechanics, chemistry, and physics, where solutions may

*Received by the editors November 18, 2024; accepted for publication (in revised form) September 19, 2025; published electronically November 19, 2025.
https://doi.org/10.1137/24M1710905

Funding: The first author's research was supported in part by the National Science Foundation under grants DMS-1912816 and DMS-2309376. The second author's research was supported in part by the National Center for Theoretical Sciences of Taiwan and the National Science and Technology Council through grant 112-2628-M-006-009-MY4.

†Department of Mathematics, North Carolina State University, Raleigh, NC 27695-8205 USA (chu@math.ncsu.edu).

‡Department of Mathematics, National Cheng Kung University, Tainan, 701 Taiwan (mhlin@mail.ncku.edu.tw).

linger near unstable equilibria for extended periods before eventually transitioning, highlighting the intricate interplay between reaction and diffusion. Understanding these transition events, including their locations in the configuration space and frequencies of these incidents, provides critical information in important applications such as phase transitions in nucleation, conformational changes in macromolecules, and transition states in chemical reactions [20]. Beyond these practical implications, saddle points can be viewed as intermediate states of specific dynamical flows that separate distinct phases or behaviors, highlighting their critical role in theoretical and applied contexts [4]. As a result, computing a saddle point is essential for many important applications. Various approaches and some sophisticated numerical techniques have been developed. See [4, 9, 11, 15, 16, 38, 42, 44], to mention a few.

One plausible idea for finding the saddle point is to search by a smooth evolution [21]. Given an energy function $f:\mathbb{R}^n \to \mathbb{R}$, an elegant formulation first introduced in [17] is the so-called gentlest ascent dynamics (GAD), which orchestrates the evolution of both a position vector $\mathbf{x}$ and a direction vector $\mathbf{v}$ through the coupled system:

$$\begin{cases} \dot{\mathbf{x}} = -\nabla f(\mathbf{x}) + 2\langle \nabla f(\mathbf{x}), \mathbf{v}\rangle \mathbf{v}, \\ \dot{\mathbf{v}} = -\nabla^2 f(\mathbf{x})\mathbf{v} + \langle \mathbf{v}, \nabla^2 f(\mathbf{x})\mathbf{v}\rangle \mathbf{v}, \quad \|\mathbf{v}(0)\| = 1, \end{cases} \tag{1.1}$$

where ∇ and ∇^2 stand for the gradient and the Hessian operators, respectively. The GAD is claimed to be capable of escaping a basin of attraction (of the gradient of f). It has been shown that if the flow ever converges, then the limit point $(\mathbf{x}^*, \mathbf{v}^*)$ of (1.1) must be such that $\mathbf{x}^*$ is an index-1 saddle point of f and $\mathbf{v}^*$ points to the unstable (ascent) direction at this saddle point, respectively [17, p. 1833]. This interesting dynamical system has inspired additional research efforts on multiple fronts, including its discrete analogues [6, 18], generalization to constrained manifolds [31] or adaptive point-clouds [3], and generalization to higher-index problems [32, 40, 51, 52].

The main contribution of this paper is to shed new insight on the working mechanism behind (1.1), allowing us to make the connection to several existing flows, e.g., the Oja flow [37, 50] and the double bracket flow [8, 13, 14], which were designed originally for other purposes. One important consequence of this understanding is a natural generalization of (1.1) to dynamical systems effective for high-index problems. Another consequence of this connection is that the proof of convergence follows from what we have already known about these existing flows.

This paper is organized as follows. We begin in section 2 with a careful interpretation of the geometry involved in (1.1). We identify the criterion that determines whether the objective value increases or decreases. We generalize this notion to higher-index problems in section 3 by tracking certain subspaces in the Grassmann manifold. One approach is to convert the Oja flow to our advantage, where we track the orthonormal basis for a subspace flow. We argue that it is a natural generalization of the continuous power method. Another equivalent but more convenient approach is to track the corresponding orthogonal projector, which leads to a matrix Riccati differential equation. We further convert it to a double bracket equation which enjoys a sorting property, whence the stability analysis is known. As far as eigenspace calculation is concerned, perhaps the most straightforward way is the conventional subspace iteration with QR normalization. This approach leads to a hybrid method. The convergence analysis is presented in section 4, where we first classify all possible equilibrium points for each dynamical system, and then we establish the theory of stability with different properties. In section 5, we turn our attention to the numerical resolution of the differential systems. After briefly surveying several representative

discrete schemes for GAD, we focus on the use of high-precision numerical integrators. This approach enables us to clearly reveal the distinctive dynamical features inherent to the underlying differential systems through a series of carefully designed numerical experiments.

2. Basic mechanism. It is important to understand the workings behind (1.1) so that we can generalize it properly to the higher-index case.[1] The motivation has been mentioned as "intuitive" in [17], but we elaborate the ideas further as follows.

For a fixed $\mathbf{x}$, the vector field for $\mathbf{v}$ in (1.1) is precisely the negative projected gradient of the Rayleigh quotient

$$\rho(\mathbf{v};\mathbf{x}) := \langle \mathbf{v}, \nabla^2 f(\mathbf{x})\mathbf{v}\rangle, \quad \mathbf{v} \in S^{n-1}, \tag{2.1}$$

onto the unit sphere $S^{n-1} \subset \mathbb{R}^n$. Therefore, the solution flow $\mathbf{v}(t)$ stays on S^{n-1} if $\mathbf{v}(0) \in S^{n-1}$ and its instantaneous motion points to the direction along which $\rho(\mathbf{v};\mathbf{x})$ is decreasing. With this fixed $\mathbf{x}$, it can be verified that $\mathbf{v}(t)$ has a closed-form solution

$$\mathbf{v}(t) = \frac{e^{-\nabla^2 f(\mathbf{x})t}\mathbf{v}_0}{\|e^{-\nabla^2 f(\mathbf{x})t}\mathbf{v}_0\|}. \tag{2.2}$$

The projected gradient flow for $\mathbf{v}(t)$ has thus been known as the continuous power method in [12]. It follows that $\mathbf{v}(t)$ will converge to the unit eigenvector of $\nabla^2 f(\mathbf{x})$ associated with its leftmost eigenvalue. For a saddle point $\mathbf{x}^*$ of index 1, $\nabla^2 f(\mathbf{x}^*)$ has exactly one negative eigenvalue. The associated eigenvector $\mathbf{v}^*$ is referred to as the *unstable* direction. By continuity, we know that if $\mathbf{x}(t)$ is near $\mathbf{x}^*$, then $\mathbf{v}(t)$ should move to approximately align with the unstable direction $\mathbf{v}^*$. When this happens, we can conceptualize our position $\mathbf{x}(t)$ as being on the "concave-down" side of the surface defined by $G := \{(\mathbf{x}, z) \in \mathbb{R}^n \times \mathbb{R} | z = f(\mathbf{x})\}$. Thus, it is natural to require that $f(\mathbf{x}(t))$ should be increasing when $\mathbf{x}(t)$ climbing up to $\mathbf{x}^*$ along this direction. To fulfill this requirement, we vary $\mathbf{x}(t)$ according to the first differential equation in (1.1), which itself has a remarkable functionality that we elucidate below.

The vector field for $\mathbf{x}$ in (1.1) is actually the Householder reflector $W_{\mathbf{v}} : \mathbb{R}^n \to \mathbb{R}^n$,

$$W_{\mathbf{v}} := I_n - 2\mathbf{v}\mathbf{v}^\top, \tag{2.3}$$

applied to the vector $-\nabla f(\mathbf{x})$. Recall that $W_{\mathbf{v}}$ is an orthogonal transformation with the action that $W_{\mathbf{v}}\mathbf{z}$ of any $\mathbf{z} \in \mathbb{R}^n$ is simply the mirror image of $\mathbf{z}$ with respect to the hyperplane normal to the unit vector $\mathbf{v}$. When $\mathbf{v}(t)$ is gradually aligned with the unstable direction $\mathbf{v}^*$, we want $\mathbf{x}(t)$ to move toward a direction to increase the objective value $f(\mathbf{x})$. In theory, we could use the gradient direction for the steepest ascent, but that would disengage the information of $\mathbf{v}$ from affecting the movement of $\mathbf{x}(t)$ continuously. The vector field in (1.1) has the effect of turning $-\nabla f(\mathbf{x})$ by an angle through the Householder reflector with the hope that $\dot{\mathbf{x}}$ points to a direction of ascent, but this can happen if and only if

$$\langle -\nabla f(\mathbf{x}) + 2\langle \nabla f(\mathbf{x}), \mathbf{v}\rangle \mathbf{v}, \nabla f(\mathbf{x})\rangle > 0. \tag{2.4}$$

We quickly deduce that (2.4) holds if and only if the angle between $\nabla f(\mathbf{x})$ and $\mathbf{v}$ or $-\mathbf{v}$ is less than $\frac{\pi}{4}$. If the angle is larger than $\frac{\pi}{4}$, we may think that $\nabla f(\mathbf{x})$ is about

[1] The Morse index of a critical point refers to the number of negative eigenvalues of the Hessian at that point. Therefore, a local minimum is of index 0 (all eigenvalues positive), and a local maximum is of index n (all eigenvalues negative). A critical point with index k in between is called an index-k saddle point [34, 52].

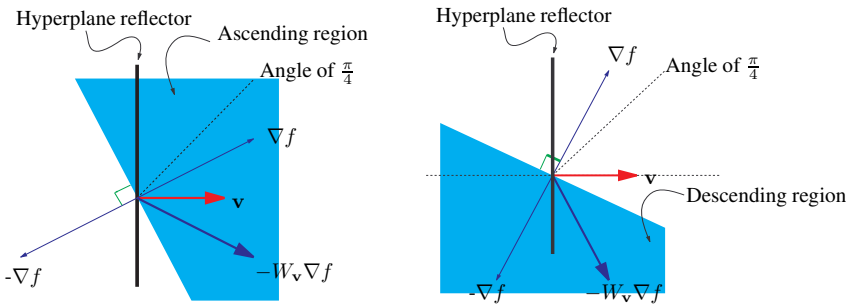

FIG. 1. *Basic mechanism in the GAD.*

orthogonal to $\mathbf{v}$, but by construction we know that $\mathbf{v}(t)$ is more or less aligned with the unstable direction. This suggests that the position $\mathbf{x}$ might be on the "concave-up" side of the surface G. In that case, it is natural to require that $f(\mathbf{x})$ decrease. In short, this critical angle $\frac{\pi}{4}$ is used as the threshold to decide whether the flow should ascend or descend. In either decision, the Householder reflection applied to $-\nabla f(\mathbf{x})$ as proposed in (1.1) is providing the desired mechanism automatically. We depict the mechanism in Figure 1.

We do need to point out that the abovementioned reasoning is only intuitively true. Its success lies in the fact that the location $\mathbf{x}(t)$ and the vector $\mathbf{v}(t)$ must work in tandem, which depends on the initial values. To illustrate our point, consider the simple problem $f(x_1, x_2) = x_1^2 - x_2^2$. Its Hessian $\nabla^2 f(\mathbf{x}) = \begin{bmatrix} 2 & 0 \\ 0 & -2 \end{bmatrix}$ is totally independent of $\mathbf{x}$. The vector $\mathbf{v}(t)$ stays on the unit circle S^1, evolves by its own dynamics, and is not influenced by $\mathbf{x}(t)$. If we write $\mathbf{v}(t) = [\cos\theta(t), \sin\theta(t)]^\top$, then it is easy to see that

$$\dot{\theta} = 2\sin 2\theta,$$

which manifests how $\mathbf{v}(t)$ rotates on S^1. Specifically, starting from any point on the upper semicircle the flow $\mathbf{v}(t)$ moves toward the equilibrium $[0,1]^\top$, whereas starting from points on the lower semicircle $\mathbf{v}(t)$ converges to $[0,-1]^\top$. Both of these equilibrium points are the unstable directions of $\nabla^2 f(\mathbf{x})$ for any $\mathbf{x}$. Consider the scenario that $\mathbf{x}_0 = [1, \frac{1}{2}]^\top$, so $-\nabla f(\mathbf{x}_0) = [-2, 1]^\top$. At this position $f(\mathbf{x}_0) > 0$, so we would like to see $\mathbf{x}(t)$ to move toward the origin to decrease the objective value. However, if $\mathbf{v}_0$ is chosen such that it points to the angle $\theta_0 \in (\arctan(\frac{-1}{2}) - \frac{\pi}{4}, \arctan(\frac{-1}{2}) + \frac{\pi}{4})$, then $f(\mathbf{x}(t))$ actually increases initially until eventually the rotation of $\mathbf{v}(t)$ causes $\mathbf{x}(t)$ to turn around and move toward the origin.

3. Generalization to high-index saddle points. The GAD is but one of many numerical algorithms proposed for computing index-1 saddle point problems which arise in many practical problems. While most existing algorithms for solving index-1 saddle point problems have been well designed and studied, the development of effective methods for computing high-index saddle points has been slow despite the importance of high-index problems. This difficulty is partly because of the need to estimate the high-index k in advance, and partly the need to determine the descent (unstable) directions. The method HiOSD proposed in [51], for example, estimates the k unstable eigenvectors by minimizing k Rayleigh quotients simultaneously subject

to the min-max principle imposed by the Courant–Fischer theorem. By means of the method of Lagrangian multiplier, the GAD has been generalized to the dynamical system

(3.1)
$$\begin{cases} \dot{\mathbf{x}} = -\nabla f(\mathbf{x}) + 2\sum_{i=1}^{k} \langle \nabla f(\mathbf{x}), \mathbf{v}_i \rangle \mathbf{v}_i, \\ \dot{\mathbf{v}}_i = -\nabla^2 f(\mathbf{x})\mathbf{v}_i + \langle \mathbf{v}_i, \nabla^2 f(\mathbf{x})\mathbf{v}_i \rangle \mathbf{v}_i + 2\sum_{j=1}^{i-1} \langle \mathbf{v}_j, \nabla^2 f(\mathbf{x})\mathbf{v}_i \rangle \mathbf{v}_j, \qquad i = 1, \ldots, k. \end{cases}$$

In theory, the orthonormal condition $\langle \mathbf{v}_i, \mathbf{v}_j \rangle = \delta_{ij}$ is maintained in (3.1). In practice, however, it is difficult to meet this constraint. It is very likely that the orthogonality will be lost due to the propagation of integration errors. Instead of approximating the k unstable eigenvectors individually, perhaps it suffices to approximate the k-dimensional unstable subspace, which is the purpose of this paper. To achieve this goal, we outline three interesting approaches together with links to some other flows in the literature in this section. For quick reference, the ultimate schemes are given in (4.1), (4.2), and (4.3), followed by the analysis of their equilibrium points and limiting behaviors, respectively.

3.1. Continuous-time subspace flow. The first approach is the notion of continuous-time subspace flow. Let $\mathrm{span}(Y) \subset \mathbb{R}^n$ denote the columns space of the matrix $Y \in \mathbb{R}^{n\times k}$. The basic idea of subspace flow is to require that if $Y_1(t)$ and $Y_2(t)$ are solutions of the same matrix differential equation with $\mathrm{span}(Y_1(0)) = \mathrm{span}(Y_2(0))$, then $\mathrm{span}(Y_1(t)) = \mathrm{span}(Y_2(t))$ for all t in the interval of existence. See [1] for some general discourse on this subject.

Of particular interest is the Oja flow defined by the differential equation

$$\dot{X} = (I_n - XX^\top)CX, \quad X(0) = X_0, \tag{3.2}$$

where $X(t) \in \mathbb{R}^{n\times k}$ and C is a fixed matrix in $\mathbb{R}^{n\times n}$ [37, 50]. Some important properties of the solution flow $X(t)$ are summarized as follows. Proofs and more detailed discussions can be found in [50].

THEOREM 3.1. *Suppose that the matrix C in* (3.2) *is symmetric and positive definite with k dominant eigenvalues. Let columns of the matrix Υ represent the corresponding dominant eigenvectors.*[2]

1. *The solution to the initial value problem* (3.2) *is given by*
$$X(t) = e^{Ct}X_0(I_k - X_0^\top X_0 + X_0^\top e^{2Ct}X_0)^{-\frac{1}{2}}. \tag{3.3}$$
2. *The solution $X(t)$ converges exponentially to an equilibrium of* (3.2) *as $t \to \infty$.*
3. *The subspace* $\mathrm{span}(X(t))$ *converges to* $\mathrm{span}(\Upsilon)$ *if and only if*
$$\mathrm{rank}(X_0^\top \Upsilon) = k. \tag{3.4}$$
Therefore, $\mathrm{span}(X(t))$ *converges to the k-dimensional dominant subspace of C for almost all initial values X_0.*

[2]By a k-dimensional eigenspace, we mean the linear subspace spanned by k linearly independent eigenvectors. Here, because of the special choice of Υ, the subspace $\mathrm{span}(\Upsilon)$ is referred to as the k-dimension dominant eigenspace of C.

4. *If X_0 is of full column rank, let $\alpha := \min\{1, \sigma_k\}$, where σ_k denote the smallest nonzero singular value of X_0. Then*

$$\|I_k - X(t)^\top X(t)\|_F \le \|I_k - X_0^\top X_0\|_F e^{-2\alpha^2 \lambda_{min}(C)t} \quad \text{for all } t \ge 0. \tag{3.5}$$

   In particular, $\lim_{t\to\infty} X(t)^\top X(t) = I_k$.
5. *If $X_0^\top X_0 = I_k$, then $X(t)^\top X(t) = I_k$ for all $t > 0$.*
6. *If $R(X) := \mathrm{trace}(X^\top C X)$, then $R(X(t))$ is strictly increasing.*

For our application, suppose that the saddle point $\mathbf{x}^*$ is known a priori to be of index k. Then with c sufficiently large, e.g., $c = c(\mathbf{x}) > \lambda_{max}(\nabla^2 f(\mathbf{x}))$, the matrix

$$C = C(\mathbf{x}) := cI_n - \nabla^2 f(\mathbf{x}) \tag{3.6}$$

is symmetric and positive definite in a neighborhood of $\mathbf{x}^*$. Obviously, the original k negative eigenvalues of $\nabla^2 f(\mathbf{x})$ are now transformed to the k dominant eigenvalues of C, while preserving the corresponding eigenvectors. We therefore propose the dynamical system

$$\begin{cases} \dot{\mathbf{x}} = -\nabla f(\mathbf{x}) + 2\mathscr{P}_U \nabla f(\mathbf{x}), \\ \dot{U} = (I_n - UU^\top)(cI_n - \nabla^2 f(\mathbf{x}))U, \quad U(0) = U_0, \end{cases} \tag{3.7}$$

where, in the spirit of (2.3), we have applied the generalized reflector

$$W_U := I_n - 2\mathscr{P}_U \tag{3.8}$$

to the vector $-\nabla f(\mathbf{x})$, where $\mathscr{P}_U$ denotes the orthogonal projector onto the subspace $\mathrm{span}(U)$. For a fixed $\mathbf{x}$ near the index-k saddle point $\mathbf{x}^*$, the corresponding Oja flow $U(t)$ will thus converge to a limit point whose column space is the unstable eigenspace of $\nabla^2 f(\mathbf{x})$. Of course, since $\mathbf{x}(t)$ in (3.7) is varying in t, we need more rigorous convergence analysis for the coupled system (3.7) than merely an intuitive reasoning.

According to Theorem 3.1, if the initial value U_0 for (3.7) originates from the Stiefel manifold

$$\mathcal{S}(n,k) := \{X \in \mathbb{R}^{n\times k} | X^\top X = I_k\}, \tag{3.9}$$

then the solution flow $U(t)$ stays on $\mathcal{S}(n,k)$ for all $t > 0$. We will assume this condition holds throughout our discussion. Consequently, the shift c employed in (3.6) becomes irrelevant, as the identity

$$(I_n - UU^\top)(cI_n)U = 0$$

is trivially satisfied for any c. The differential equation for $U(t)$ in (3.7) is exactly the same as

$$\dot{U} = -\nabla^2 f(\mathbf{x})U + U(U^\top \nabla^2 f(\mathbf{x})U). \tag{3.10}$$

For (3.7) and, hence, for (3.10), we still can use Theorem 3.1 to draw the following results of which a remarkable fact is that we only need to focus on the k leftmost eigenvalues of $\nabla^2 f(\mathbf{x})$, irrespective of whether they are negative or not.

LEMMA 3.2. *For a fixed* $\mathbf{x}$*, let* $U(t)$ *be the solution to the differential system* (3.10) *(equally to* (3.7)*) with initial value* $U_0 \in \mathcal{S}(n,k)$*. Then*

1. *the number of positive eigenvalues of* $\nabla^2 f(\mathbf{x})$ *does not play a role at the convergence of* $U(t)$*;*
2. *suppose that the* k *leftmost eigenvalues of* $\nabla^2 f(\mathbf{x})$ *are separate from other eigenvalues. Then the subspace flow* $\mathrm{span}(U(t))$ *converges to the eigenspace corresponding to these eigenvalues from almost all* U_0*.*

The system (3.10) turns out to be the original Oja dynamics applied to $-\nabla^2 f(\mathbf{x})$, even though $-\nabla^2 f(\mathbf{x})$ is not definite at all. Equivalently, we may interpret $U(t)$ as integrating the Oja dynamics (3.2) in the negative time. Therefore, the following closed-form solution is readily available.

LEMMA 3.3. *For a fixed* $\mathbf{x}$*, let* $U(t)$ *be the solution to the differential system* (3.10) *with initial value* $U_0 \in \mathcal{S}(n,k)$*. Then*

$$U(t) = e^{-\nabla^2 f(\mathbf{x})t} U_0 (U_0^\top e^{-2\nabla^2 f(\mathbf{x})t} U_0)^{-\frac{1}{2}}. \tag{3.11}$$

It is interesting to note that the Oja flow represented by (3.10) serves as the matrix counterpart to the continuous power flow described in (1.1). Additionally, (3.11) bears a resemblance to (2.2), with the distinction that $U_0 \in \mathcal{S}(n,k)$ is a matrix. Despite these resemblances, it is crucial to emphasize that Lemmas 3.2 and 3.3 are insufficient for our theoretical framework. This is because the pair $(\mathbf{x}(t), U(t))$ must concurrently satisfy the coupled system (3.7). Further convergence analysis is needed.

A comparison of (3.7) with the HiOSD system (3.1) is due. Assume $U_0 \in \mathcal{S}(n,k)$. Then the projection $\mathscr{P}_U \nabla f(\mathbf{x})$ in (3.7) has the same meaning of a Fourier expansion as that in (3.1) except that they are using different orthonormal bases. Second, the criterion (2.4) is generalized to

$$2\langle U^\top \nabla f(\mathbf{x}), U^\top \nabla f(\mathbf{x})\rangle > \langle \nabla f(\mathbf{x}), \nabla f(\mathbf{x})\rangle, \tag{3.12}$$

which is related to the angle θ between the vectors $\pm\nabla f(\mathbf{x})$ and the subspace $\mathrm{span}(U)$. In turn, it determines

$$\frac{df(\mathbf{x}(t))}{dt} = \langle -W_U \nabla f(\mathbf{x}), \nabla f(\mathbf{x})\rangle \Longrightarrow f \text{ is } \begin{cases} \text{increasing} & \text{if (3.12) holds } (\theta < \frac{\pi}{4}); \\ \text{decreasing} & \text{otherwise } (\theta > \frac{\pi}{4}). \end{cases} \tag{3.13}$$

Third, with $V := [\mathbf{v}_1, \ldots, \mathbf{v}_k]$, the differential equations $\dot{\mathbf{v}}_i(t)$, $i = 1, \ldots, k$, in the HiOSD (3.1) can be expressed in matrix form as

$$\dot{V} = -\nabla^2 f(\mathbf{x}) V + V \mathscr{L}(V^\top \nabla^2 f(\mathbf{x}) V), \tag{3.14}$$

where $\mathscr{L} := 2\mathrm{triu} - \mathrm{diag}$ denotes the operator of twice the upper triangular part minus the diagonal part of a matrix. For a fixed $\mathbf{x}$, the dynamical behavior of (3.10) is well understood [50] and, indeed, its closed form is given by (3.11), but the analysis for (3.14) is still incomplete. Despite their similarity, the most significant difference is that the HiOSD aims at approximating the dominant eigenvectors, whereas the Oja flow aims at approximating the dominant eigenspace. Finally, we mention in passing that beyond the approaches discussed thus far, a variety of alternative dynamical systems have been proposed for computing dominant eigenspaces. For instance, the work in [27] introduces a family of generalized orthogonal flows, which encompass the Oja flow as a particular case.

Along the trajectories $(\mathbf{x}(t), U(t))$, it is possible that the index of the Hessian $\nabla^2 f(\mathbf{x}(t))$ varies in t, whereas determining the index a priori is one of the difficulties. Although Lemma 3.2 implies that even if k is not the correct index, the flow $U(t)$ will still find the eigenspace associated with the k leftmost eigenvalues of $\nabla^2 f(\mathbf{x})$; it should be noted that the mechanism that we have relied on to modulate the direction $-\nabla f(\mathbf{x})$ might fail. This can be seen by considering the scenario that k is less than the true index. Suppose that the condition (3.12) holds. Thus, we wish to move $\mathbf{x}$ to increase the objective value. The Householder reflection $W_U(-\nabla f(\mathbf{x}))$ does point to a direction of ascent. This is acceptable. However, if the condition (3.12) does not hold, then the movement dictated by $W_U(-\nabla f(\mathbf{x}))$ might cause the objective values to slide further down along those uncounted unstable directions, which is not desired.

3.2. Projector flow. Recall that the collection of all linear subspaces of fixed dimension, say, k, of $\mathbb{R}^n$ forms the Grassmann manifold $\mathscr{G}(n,k)$. An element in $\mathscr{G}(n,k)$ is actually a subspace which can be identified by an orthonormal basis $U \in \mathcal{S}(n,k)$. Equivalently, or perhaps more conveniently, the subspace can be represented by the unique orthogonal projector $P := \mathscr{P}_U = UU^\top \in \mathbb{R}^{n\times n}$ onto the subspace span(U). In this way, note that P is basis independent.

Upon substitution, we may rewrite the system (3.7) as the dynamical system

$$\begin{cases} \dot{\mathbf{x}} = -\nabla f(\mathbf{x}) + 2P\nabla f(\mathbf{x}), \\ \dot{P} = 2P\nabla^2 f(\mathbf{x})P - \nabla^2 f(\mathbf{x})P - P\nabla^2 f(\mathbf{x}), \end{cases} \tag{3.15}$$

with $P(0) = U(0)U(0)^\top$ and $U(0) \in \mathcal{S}(n,k)$. The dynamical system for $P(t)$ in (3.15) is known as the matrix Riccati differential equation (MRDE). This is a classical differential system that has been studied extensively. See, for example, [28, 30] and the many references contained therein. See also [50] for a general closed-form solution of $P(t)$. Its equilibrium points are solutions to the continuous algebraic Riccati equation (known conventionally as the CARE in the literature)

$$F^\top Z + ZF - ZGZ + H = 0, \tag{3.16}$$

with $F = F^\top = -\nabla^2 f(\mathbf{x})$, $G = -2\nabla^2 f(\mathbf{x})$ and $H = 0$. Solutions to (3.16) can be found numerically by using the Schur method [29] or the generalized eigenvalue algorithm [2].

Though mathematically equivalent, the dynamical system (3.15) for $(\mathbf{x}, P(t))$ is much harder to integrate numerically than the system (3.7) for $(\mathbf{x}, U(t))$. The reason is that a projector must satisfy the identity

$$P^2 = P, \tag{3.17}$$

but the MRDE (3.15) does not maintain such a property per se for its general trajectories [28]. Put differently, the product $U(t)U(t)^\top$ of a solution $U(t)$ to (3.7) does satisfy the MRDE, but a general solution to the MRDE is not necessarily factorizable as a product $U(t)U(t)^\top$ with $U(t)$ satisfying (3.7). Any round-off error incurred in numerical integration, therefore, will eventually cause $P(t)$ to drift away from its role as a projector. In contrast, the property (3.5) in Theorem 3.1 implies that the Stiefel manifold $\mathcal{S}(n,k)$ is an asymptotically stable manifold for the Oja dynamical system (3.2). Even if $X(t)$ is off the manifold, the contraction inherent in the dynamics will bring it back to $\mathcal{S}(n,k)$ in the long run. Our numerical experiments confirm this drifting of $P(t)$ and the contraction of $X(t)$.

To enforce the condition (3.17), we do have another way to describe the dynamical system for $P(t)$, which can circumvent the numerical drift mentioned above and maintain $P(t)$ as a projector.

LEMMA 3.4. *Assume $U_0 \in \mathcal{S}(n,k)$. For a fixed $\mathbf{x}$, let $U(t)$ be a solution to the differential system* (3.10). *Then, the product $P(t) = U(t)U(t)^\top$ satisfies the double bracket equation*

$$\dot{P} = [P, [P, -\nabla^2 f(\mathbf{x})]], \tag{3.18}$$

where $[A, B] := AB - BA$ denotes the Lie bracket.

Proof. Since $U(t) \in \mathscr{S}(n,r)$, we know in theory that $P^2(t) = P(t)$ for all t. Rewrite the differential equation in (3.15) as

$$\dot{P} = 2P\nabla^2 f(\mathbf{x})P - \nabla^2 f(\mathbf{x})P^2 - P^2\nabla^2 f(\mathbf{x}),$$

which is exactly the expansion of the double bracket in (3.18). ∎

The double bracket system (3.18) has been studied thoroughly in [5, 8, 10, 14, 22, 23, 45] with some special meanings which we summarize below. The proof follows from the general results already developed in [14].

LEMMA 3.5. *Under the same assumptions as in Lemma* 3.4. *The solution $P(t)$ to* (3.18) *has the following properties:*

1. *The flow defined by* (3.18) *is the negative projected gradient flow of the objective function*

$$F(P) := \frac{1}{2}\|P + \nabla^2 f(\mathbf{x})\|_F^2 \tag{3.19}$$

   subject to the constraint that $P \in \mathcal{M}(n,k)$ where

$$\mathcal{M}(n,k) := \{QQ^\top | Q \in \mathcal{S}(n,k)\}.$$

2. *A necessary condition for P to be a stationary point of $F(P)$ is $[P, \nabla^2 f(\mathbf{x})] = 0$.*
3. *A stationary point $P = QQ^\top \in \mathcal{M}(n,k)$ is a minimizer of $F(P)$ if and only if columns of Q form an orthonormal basis for the eigenspace corresponding to the k leftmost eigenvalues of $\nabla^2 f(\mathbf{x})$.*

The last statement in Lemma 3.5 happens to provide an alternative proof of Lemma 3.2. The reason that the minimizer must pick up an eigenspace corresponding to the k leftmost eigenvalues of $\nabla^2 f(\mathbf{x})$ is that the double bracket equation has an innate sorting property which has been exploited for other applications in [8, 14].

3.3. Subspace iteration with QR normalization. The idea behind the three coupled dynamical systems, (3.1), (3.7), and (3.15), is almost identical. The fundamental mechanism is to estimate the unstable eigenspace by the second equation and use it as a reference point to gauge how the negative gradient direction should be modified via the Householder reflector in the first equation. The reflection has the effect that if the gradient is approximately aligned with the unstable eigenspace in the sense of (3.12), then $\mathbf{x}(t)$ moves in the ascending direction; otherwise, it moves in the descending direction. The adaption of both the eigenspace and the position vector is continuous, simultaneous, but infinitesimal. However, path following is slow,

expensive, and prone to error propagation in computation. The linear algebra might offer a cheap alternative way to alleviate the situation and still meet the purpose.

Subspace iteration, a rather straightforward generalization of the classical single-vector power iteration, can be used to compute eigenspaces of matrices [43]. Suppose that $C \in \mathbb{R}^{n\times n}$ is a symmetric and positive definite matrix with k dominant eigenvalues. Starting with an arbitrary rank-k matrix $Q_0 \in \mathbb{R}^{n\times k}$, it is know that the iteration

(3.20)
$$\begin{cases} Y_{k+1} := CQ_k, & k = 0, 1, 2, \ldots, \\ Q_{k+1} := \text{ the } Q \text{ factor in the compact } QR \text{ decomposition } Y_{k+1} = Q_{k+1}R_{k+1}, \end{cases}$$

converges to an orthonormal basis for the eigenspace of the k dominant eigenvalues [47]. Denote

$$\mathscr{I}(C; Q_0) := \lim_{k\to\infty} Q_k. \tag{3.21}$$

By requiring that the R factor in the QR decomposition to always maintain positive diagonal entries, it can be argued that $\mathscr{I}(C; Q_0)$ is continuous in C [25].

Motivated by the discussion in the preceding sections where finding the dominant eigenspace is the central issue, we may introduce the fourth dynamical system

$$\begin{cases} \dot{\mathbf{x}} = -\nabla f(\mathbf{x}(t)) + 2Q(t)Q(t)^\top \nabla f(\mathbf{x}(t)), \\ Q(t) = \mathscr{I}(C(\mathbf{x}(t)); Q_0)), \end{cases} \tag{3.22}$$

where

$$C(\mathbf{x}(t)) := cI_n - \nabla^2 f(\mathbf{x}(t)) \tag{3.23}$$

with $c = c(\mathbf{x}(\mathbf{t}))$ large enough so that $C(\mathbf{x}(t))$ is positive definite. The system (3.22) makes sense in theory except that the value of the continuous function $Q(t)$ is obtained from the subspace iteration scheme (3.20) per given t. On the other hand, in view of the fact that $U(t)$ and $P(t)$ produced by the systems (3.7) and (3.15) only approximate the eigenspace, respectively, there is no need to calculate $Q(t)$ to high precision per given t. That is, there is no need to take the limit in (3.21). Perhaps it suffices to take just a finite number of iterations in (3.20) to create a reflector in (3.22) to influence the behavior of the flow $\mathbf{x}(t)$. In this case, we may interpret $\mathscr{I}$ as a discrete dynamical system.

4. Convergence analysis. For clarity, we recapitulate our own three dynamical systems for the saddle point calculation as follows:

1. (Continuous subspace flow)

$$\begin{cases} \dot{\mathbf{x}} = -\nabla f(\mathbf{x}) + 2UU^\top \nabla f(\mathbf{x}), \\ \dot{U} = -\nabla^2 f(\mathbf{x})U + U(U^\top \nabla^2 f(\mathbf{x})U), \quad U(0) = U_0 \in \mathcal{S}(n,k). \end{cases} \tag{4.1}$$

2. (Projector flow)

$$\begin{cases} \dot{\mathbf{x}} = -\nabla f(\mathbf{x}) + 2P\nabla f(\mathbf{x}), \\ \dot{P} = [P, [P, -\nabla^2 f(\mathbf{x})]], \quad P(0) = U_0U_0^\top. \end{cases} \tag{4.2}$$

3. (Discrete subspace iteration)

$$\begin{cases} \dot{\mathbf{x}} = -\nabla f(\mathbf{x}(t)) + 2Q(t)Q(t)^\top \nabla f(\mathbf{x}(t)), \\ Q(t) = \mathscr{I}(C(\mathbf{x}(t)); Q_0)), \quad C(\mathbf{x}(t)) = c(t)I_n - \nabla^2 f(\mathbf{x}(t)). \end{cases} \tag{4.3}$$

In this section, we identify the equilibrium points of the dynamical system and conduct a local stability analysis in their vicinity. For related discussions, we refer the reader to [32] for error estimates concerning the Euler discretization of (3.1), as well as [18, 52] for results on other discrete schemes. Our analysis is confined to local properties; a complete characterization of the global dynamics of these systems remains a challenging and largely open problem.

4.1. Equilibrium points. In this section we characterize the equilibrium points for each of the dynamical systems without considering their stability.

THEOREM 4.1. *Let* $(\mathbf{x}^*, X^*) \in \mathbb{R}^n \times \mathbb{R}^{n\times k}$ *denote an equilibrium point of any of the systems* (4.1), (4.2), *and* (4.3), *where* X^* *is interpreted differently according to the specific system.*

1. $\mathbf{x}^*$ *must be a critical point of the energy function* $f(\mathbf{x})$, *i.e.,* $\nabla f(\mathbf{x}^*) = 0$.
2. *Let* $(\lambda_i, \mathbf{v}_i)$, $i = 1, \ldots, n$, *denote the orthonormal eigenpairs of the Hessian matrix* $\nabla^2 f(\mathbf{x}^*)$. *Assume that all eigenvalues are distinct and* $\lambda_1 < \lambda_2 < \cdots < \lambda_n$. *Then, the equilibrium point* X^* *for each of the systems can be described as follows:*
   (a) *For the system* (4.1), *columns of* X^* *form an orthonormal basis of a* k*-dimensional eigenspace of* $\nabla^2 f(\mathbf{x}^*)$. *There are* $\binom{n}{k}$ *distinct* k*-dimensional eigenspaces, but infinitely many ways to represent the basis.*
   (b) *For the system* (4.2), X^* *must be the projector operator of a* k*-dimensional eigenspace of* $\nabla^2 f(\mathbf{x}^*)$. *There are* $\binom{n}{k}$ *distinct* k*-dimensional projectors.*
   (c) *For the system* (4.3), $X^* = [\mathbf{v}_1, \ldots, \mathbf{v}_k]D$, *where* D *is a diagonal matrix with* ± 1 *along the diagonal. There are* 2^k *distinct equilibrium points.*

Proof. First, all three schemes, (4.1), (4.2), and (4.3), resort to a projector $\mathscr{P}$ defined by $UU^\top$, P, or $QQ^\top$ with $U, Q \in \mathcal{S}(n,k)$, and $P^2 = P$, respectively. Consequently, the corresponding Householder reflector is orthogonal. Thus, if $(\mathbf{x}^*, X^*) \in \mathbb{R}^n \times \mathbb{R}^{n\times k}$ is an equilibrium point of any of these systems, it follows necessarily that $\nabla f(\mathbf{x}^*) = 0$.

For clarity, we differentiate the equilibrium point X^* for systems (4.1), (4.2), and (4.3) by the notation U^*, P^*, and Q^*, respectively. It is convenient to let

$$\nabla^2 f(\mathbf{x}^*) = V \Lambda V^\top \tag{4.4}$$

denote the spectral decomposition of $\nabla^2 f(\mathbf{x}^*)$. We analyze their structures separately.

If $(\mathbf{x}^*, U^*)$ is an equilibrium point of (4.1), then it must satisfy the equation

$$\nabla^2 f(\mathbf{x}^*) U^* = U^* ({U^*}^\top \nabla^2 f(\mathbf{x}^*) U^*), \tag{4.5}$$

which implies that $\operatorname{span}(U^*)$ is invariant under the transformation $\nabla^2 f(\mathbf{x}^*)$. Suppose that $U^* \in \mathcal{S}(n,k)$ is a solution to (4.5). We may express

$$U^* = VD$$

with some $D \in \mathcal{S}(n,k)$. It follows that D must satisfy the equation

$$\Lambda D = D(D^\top \Lambda D). \tag{4.6}$$

That is, D is invariant under the diagonal matrix Λ. It is known that the invariant subspace of a diagonal matrix with distinct diagonal entries is the direct sum of some subset of its 1-dimensional eigenspaces. Therefore, $D \in \mathbb{R}^{n\times k}$ must be such that

$$\operatorname{span}(D) = \operatorname{span}([\mathbf{e}_{i_1}, \ldots, \mathbf{e}_{i_k}]), \tag{4.7}$$

where $\mathbf{e}_i$, $i = 1, \ldots, n$, are the standard unit vectors. If

$$\widetilde{V} := [\mathbf{v}_{i_1}, \ldots, \mathbf{v}_{i_k}] \tag{4.8}$$

denotes a collection of k orthonormal eigenvectors, then

$$U^* = \widetilde{V}\Psi, \tag{4.9}$$

where $\Psi \in \mathbb{R}^{k\times k}$ is an arbitrary orthogonal matrix. We have thus proved that $\mathrm{span}(U^*)$ is an eigenspace of $\nabla^2 f(\mathbf{x}^*)$. There are $\binom{n}{k}$ many ways to choose k-dimensional eigenspaces.

The necessary condition for P^* to be an equilibrium point for (4.2) is that [14, Lemma 3.1]

$$[P^*, \nabla^2 f(\mathbf{x}^*)] = 0.$$

It is known that diagonalizable matrices that commute share a common basis of eigenvectors. If we write the projector as $P^* = Q^* Q^{*\top}$ with $Q^* \in \mathcal{S}(n,k)$, then columns of Q^* are eigenvectors of P^* associated with eigenvalue 1. Since we have assumed that $\nabla^2 f(\mathbf{x}^*)$ has simple eigenvalues, the columns of Q^* must be eigenvectors of the form (4.8). We have thus proved that P^* must be the projector operator of a k-dimensional eigenspace.

If $(\mathbf{x}^*, Q^*)$ is an equilibrium point of (4.3), then it is necessary that

$$C(\mathbf{x}^*)Q^* = Q^* R^*. \tag{4.10}$$

Since the upper triangular matrix $R^* = Q^{*\top} C(\mathbf{x}^*) Q^*$ must be symmetric, it is the diagonal matrix of dominant eigenvalues of $C(\mathbf{x}^*)$. Therefore, the columns of Q^* are the orthonormal eigenvectors associated with the k leftmost eigenvalues of $\nabla^2 f(\mathbf{x}^*)$. ∎

There is a quick way to see (4.7). By regarding (4.6) as a (nonlinear) eigenvalue problem

$$(D^\top \Lambda D) D^\top = D^\top \Lambda, \tag{4.11}$$

we see that the nonzero columns of $D^\top$ are eigenvectors of the $k \times k$ matrix $D^\top \Lambda D$. If Λ has distinct diagonal entries, then $D^\top$ must have $n-k$ zero columns. The indices $\{i_1, \ldots, i_k\}$ of the nonzero columns point to the nonzero rows of D. That is, columns of D are spanned by $\{\mathbf{e}_{i_1}, \ldots, \mathbf{e}_{i_k}\}$. The equality of (4.7) follows from the fact that D is rank k.

In the above, the assumption that all eigenvalues are distinct is generic, but is not necessary. In the event that $\nabla^2 f(\mathbf{x}^*)$ has multiple eigenvalues, some modified results can still be obtained. For example, the same result holds for the system (4.1), so long as we have $\lambda_1 \le \cdots \le \lambda_k < \lambda_{k+1} \le \cdots \le \lambda_n$. Consider the scenario that $k = 2$, $\lambda_1 = \lambda_2$, and the remaining $n-2$ eigenvalues are all distinct for the system (4.1). Then for (4.6), the solution $D^\top$ in (4.11) is allowed to have three nonzero columns, so long as two are associated with $\lambda_1 = \lambda_2$. Furthermore, in addition to invariant subspaces in the basic forms such as $\mathrm{span}([\mathbf{e}_{i_1}, \mathbf{e}_{i_j}])$ or $\mathrm{span}([\mathbf{e}_1 + \mathbf{e}_2, \mathbf{e}_{i_j}])$, we also have $\mathrm{span}(\Theta[\mathbf{e}_{i_1}, \mathbf{e}_{i_2}])$ as invariant subspaces, where $\Theta \in \mathbb{R}^{n\times n}$ is a block diagonal matrix with a 2×2 orthogonal block at the upper left corner and 1×1 blocks with value ± 1 along the remaining diagonal. The key point is that Θ commutes with Λ. This argument can be extended to the following result, which we stress is only a sufficient condition. We can construct counterexamples which are contrary to the necessity wrongly claimed in [50].

COROLLARY 4.2. *Let* $\widetilde{V} \in \mathbb{R}^{n\times k}$ *be given as in* (4.8). *If* $\Theta \in \mathbb{R}^{n\times n}$, $\Psi \in \mathbb{R}^{k\times k}$ *are arbitrary orthogonal matrices with* $[\Theta, \nabla^2 f(\mathbf{x}^*)] = 0$, *then the matrix*

$$X^* = \Theta \widetilde{V} \Psi \tag{4.12}$$

is an equilibrium point for the system (4.1).

Proof. We have already argued that $U^* = W\Psi$ is a solution. If $[\Theta, \nabla^2 f(\mathbf{x}^*] = 0$, it is straightforward to see by substitution that X^* in the form (4.12) also satisfies (4.5). ∎

4.2. Stability of $(\boldsymbol{x^*, X^*})$. In the above we have only characterized the equilibrium points $(\mathbf{x}^*, X^*)$ for each of the dynamical systems. Apparently, there could be many equilibrium points. We are interested in only the few which are asymptotically stable under the corresponding dynamical system.

To facilitate the discussion, we begin with the analysis by fixing one variable at a time. We first consider the stability of X^* with fixed $\mathbf{x}^*$. In the following lemma, the emphasis is that the flow $X(t)$ corresponding to each of the second differential equations in the dynamical systems (4.1), (4.2), and (4.3) converges from almost all starting values. The exceptions are those starting values satisfying some rank conditions similar to (3.4).

LEMMA 4.3. *Let* $(\mathbf{x}^*, X^*) \in \mathbb{R}^n \times \mathbb{R}^{n\times k}$ *denote an equilibrium point. Assume that the Hessian matrix* $\nabla^2 f(\mathbf{x}^*)$ *has distinct eigenvalues, as articulated in Theorem* 4.1, *and that the columns of the starting point do not span any of the* k*-dimensional eigenspaces mentioned in Theorem* 4.1, *which are themselves equilibrium points. With* $\mathbf{x}^*$ *held constant in each of the second differential equations of the systems* (4.1), (4.2), *and* (4.3), *the associated solution trajectories* $U(t)$, $P(t)$, *and* $Q(t)$ *converge globally from almost all starting values to their respective limit points. Specifically,*

1. *for the continuous subspace flow* (4.1), *the columns of* $U(t)$ *converge to an orthonormal basis for the eigenspace corresponding to the* k *leftmost eigenvalues of* $\nabla^2 f(\mathbf{x}^*)$;
2. *for the projector flow scheme* (4.2), *the differential equation is a double bracket flow that converges to the projector of the eigenspace corresponding to the* k *leftmost eigenvalues of* $\nabla^2 f(\mathbf{x}^*)$;
3. *for the discrete subspace iteration scheme* (4.3), *the limit point* $\mathscr{I}(C(\mathbf{x}^*))$, *up to some* ± 1 *sign changes in the columns, is always stable.*

Proof. For the Oja flow U(t), its limiting behavior is already characterized by Lemma 3.2, which is a consequence of Theorem 3.1 proved in [50].

The dynamical behavior of the projector flow $P(t)$ is best understood as a projected gradient flow. Its convergence to the projector of the eigenspace corresponding to the k leftmost eigenvalues of $\nabla^2 f(\mathbf{x}^*)$ is stated in Lemma 3.5, which is a consequence of a more general framework studied in [14].

The convergence of the subspace iteration (3.20) to the eigenspace of the k leftmost eigenvalues of $\nabla f(\mathbf{x}^*)$ is proved in [43, 47]. ∎

We then consider the stability of $\mathbf{x}^*$ with fixed X^*. Recall that X^* may be interpreted differently as stated in Theorem 4.1. In all three dynamical systems, X^* induces a generalized Householder reflector W in the form (3.8), whence we drop the reference to X^*, and the differential equation for $\mathbf{x}(t)$ is of the same form

$$\dot{\mathbf{x}} = -W\nabla f(\mathbf{x}). \tag{4.13}$$

The next lemma, under the index k condition, asserts the local convergence of $\mathbf{x}(t)$.

LEMMA 4.4. *Let X^* be an asymptotically stable equilibrium as described in Lemma 4.3 and W be the associated generalized Householder reflector. Assume that $\nabla^2 f(\mathbf{x}^*)$ has exactly k negative eigenvalues. Assume also that the remaining $n-k$ eigenvalues are positive.*[3] *Then, starting from a neighborhood of $\mathbf{x}^*$, the solution flow $\mathbf{x}(t)$ of the system (4.13) converges to the equilibrium $\mathbf{x}^*$, i.e., $\mathbf{x}^*$ is asymptotically stable.*

Proof. It suffices to consider the linearized system of (4.13) near $\mathbf{x}^*$. Let $\mathbf{y} := \mathbf{x} - \mathbf{x}^*$. Then, it is easy to see that

$$\dot{\mathbf{y}} = -W(\nabla^2 f(\mathbf{x}^*)\mathbf{y} + O(\|\mathbf{y}\|^2)).$$

Without loss of generality, we may rearrange the spectral decomposition (4.4) as

$$\nabla^2 f(\mathbf{x}^*) = [V_-, V_+] \begin{bmatrix} \Lambda_- & 0 \\ 0 & \Lambda_+ \end{bmatrix} \begin{bmatrix} V_-^\top \\ V_+^\top \end{bmatrix} = V_- \Lambda_- V_-^\top + V_+ \Lambda_+ V_+^\top, \tag{4.14}$$

where $\Lambda_\mp$ denote the diagonal matrices with negative and positive eigenvalues, respectively, and columns of $V_\mp$ are the corresponding orthonormal eigenvectors. Then, regardless of how the basis of X^* is chosen, the induced W is of the form

$$W = I_n - 2V_- V_-^\top.$$

It follows after simplification that

$$-W\nabla^2 f(\mathbf{x}^*) = [V_-, V_+] \begin{bmatrix} \Lambda_- & 0 \\ 0 & -\Lambda_+ \end{bmatrix} \begin{bmatrix} V_-^\top \\ V_+^\top \end{bmatrix}. \tag{4.15}$$

That is, the eigenvalues of the coefficient matrix $-W\nabla^2 f(\mathbf{x}^*)$ are all negative. ∎

The above two lemmas about one-direction stability analysis should have shed enough light on the overall stability of $(\mathbf{x}^*, X^*)$. We complete the theory by bringing in the index k condition and considering each system separately. We begin with the following general result for the system (4.1), which is of interest in its own right.

THEOREM 4.5. *Suppose that $(\mathbf{x}^*, U^*)$ is an equilibrium point for the continuous subspace dynamical system (4.1) under the conditions that $\nabla f(\mathbf{x}^*) = 0$ and U^* is an orthonormal basis for the eigenspace corresponding to the k negative eigenvalues of $\nabla^2 f(\mathbf{x}^*)$. Then, starting from a neighborhood of $(\mathbf{x}^*, U^*)$, the solution flow $(\mathbf{x}(t), U(t))$ converges to $(\mathbf{x}^*, U^*)$.*

Proof. Let the vector field in (4.1) be abbreviated as

$$\begin{aligned} F_1(\mathbf{x}, U) &:= -\nabla f(\mathbf{x}) + 2UU^\top \nabla f(\mathbf{x}), \\ F_2(\mathbf{x}, U) &:= -\nabla^2 f(\mathbf{x})U + U(U^\top \nabla^2 f(\mathbf{x})U). \end{aligned}$$

The local stability analysis depends on the eigenvalues of the (formal) Jacobian matrix

$$\mathbb{J}(\mathbf{x}, U) := \begin{bmatrix} \frac{\partial F_1(\mathbf{x},U)}{\partial \mathbf{x}} & \frac{\partial F_1(\mathbf{x},U)}{\partial \mathbf{U}} \\ \frac{\partial F_2(\mathbf{x},U)}{\partial \mathbf{x}} & \frac{\partial F_2(\mathbf{x},U)}{\partial \mathbf{U}} \end{bmatrix} \tag{4.16}$$

[3] If $\nabla^2 f(\mathbf{x}^*)$ possesses zero eigenvalues, a more nuanced analysis using center manifold theory is necessary to assess the stability of the degenerate equilibrium; this is beyond the scope of the present paper.

at the point $(\mathbf{x}^*, U^*)$. However, since the variable U is a matrix in $\mathbb{R}^{n\times k}$, it is better to interpret the partial Jacobians as linear operators via the Fréchet derivative.

Specifically, we see that the actions, denoted by the dot ".", of the partial derivatives with respect to $\mathbf{x}$ at an arbitrary $\mathbf{h}\in\mathbb{R}^n$ are given by

$$\frac{\partial F_1(\mathbf{x},U)}{\partial \mathbf{x}}.\mathbf{h} = (-I_n + 2UU^\top)\nabla^2 f(\mathbf{x}).\mathbf{h}, \tag{4.17}$$

$$\frac{\partial F_2(\mathbf{x},U)}{\partial \mathbf{x}}.\mathbf{h} = -\left(\left(\frac{\partial \nabla^2 f(\mathbf{x})}{\partial \mathbf{x}}.\mathbf{h}\right)U + U\left(U^\top\left(\frac{\partial \nabla^2 f(\mathbf{x})}{\partial \mathbf{x}}.\mathbf{h}\right)U\right)\right), \tag{4.18}$$

respectively. Note that the operator $\frac{\partial \nabla^2 f(\mathbf{x})}{\partial \mathbf{x}}$ is actually an order-3 tensor whose action on $\mathbf{h}$ to produce a matrix in $\mathbb{R}^{n\times n}$ will be complicated. In the meantime, the actions of the partial derivatives with respect to U at an arbitrary $H\in\mathbb{R}^{n\times k}$ are given by

$$\frac{\partial F_1(\mathbf{x},U)}{\partial U}.H = 2(HU^\top + UH^\top)\nabla f(\mathbf{x}), \tag{4.19}$$

$$\frac{\partial F_2(\mathbf{x},U)}{\partial U}.H = -\nabla^2 f(\mathbf{x})H + H(U^\top\nabla^2 f(\mathbf{x})U) + U(H^\top\nabla^2 f(\mathbf{x})U) + U(U^\top\nabla^2 f(\mathbf{x})H), \tag{4.20}$$

respectively. It is crucial to observe that

$$\frac{\partial F_1(\mathbf{x}^*,U^*)}{\partial U}.H = 0 \quad \text{for all } H\in\mathbb{R}^{n\times k},$$

implying that $\mathbb{J}(\mathbf{x}^*,U^*)$ is block lower triangular. Therefore, the eigenvalues of $\mathbb{J}(\mathbf{x}^*,U^*)$ are those of the two operators at the two diagonal blocks. In this way, we effectively avoid the complication of the tensor operator $\frac{\partial \nabla^2 f(\mathbf{x})}{\partial \mathbf{x}}$. The formulas (4.17) and (4.20) already express how the two operators act at "vectors" in their respective spaces. We now seek their eigenvalues. Their corresponding eigenvectors can also be characterized, but are not essential in our analysis.

For (4.17), we have already seen the spectral decomposition of the operator $\frac{\partial F_1(\mathbf{x}^*,U^*)}{\partial \mathbf{x}}$ in (4.15). In particular, its eigenvalues are the diagonal elements of Λ_- and $-\Lambda_+$ which are all negative, and its eigenvectors are precisely those of $\nabla^2 f(\mathbf{x}^*)$.

For (4.20), the operator $\frac{\partial F_2(\mathbf{x},U)}{\partial U}$ is not readily in the matrix form, but the notion of eigenvalue/eigenvector is the same, i.e., we seek all nonzero matrices H of size $n\times k$ and all scalars μ, possibly complex, such that

$$-\nabla^2 f(\mathbf{x}^*)H + H(U^{*\top}\nabla^2 f(\mathbf{x})U^*) + U^*(H^\top\nabla^2 f(\mathbf{x})U^*) + U^*(U^{*\top}\nabla^2 f(\mathbf{x})H) = \mu H. \tag{4.21}$$

By Lemma 4.3, we know that U^* is of the form $U^* = V_-\Psi$ with some orthogonal matrix $\Psi\in\mathbb{R}^{k\times k}$. The eigenvalue problem (4.21) is expected to have nk eigenvalues, counting multiplicity. Using (4.14) and making a change of variable

$$\widetilde{H} := H\Psi^\top,$$

we see that (4.21) is equivalent to

$$-V_+\Lambda_+V_+^\top\widetilde{H} + \widetilde{H}\Lambda_- + V_-\widetilde{H}^\top V_-\Lambda_- = \mu\widetilde{H}. \tag{4.22}$$

Based on (4.22), we now derive two sets of necessary conditions that characterize the eigenvalue μ. (We are not interested in the eigenvectors $\widetilde{H}$, although some partial information can be retrieved.)

First, define

$$Y := V_+^\top \widetilde{H} \in \mathbb{R}^{(n-k)\times k}.$$

Premultiply (4.22) by $V_+^\top$. We thus reduce (4.22) to

$$-\Lambda_+ Y + Y\Lambda_- = \mu Y. \tag{4.23}$$

Equivalently, for the existence of Y, it must be that

$$(-\lambda_{k+i} + \lambda_j) y_{ij} = \mu y_{ij} \tag{4.24}$$

for all $i = 1, \ldots, n-k$ and $j = 1, \ldots k$. If for a fixed pair of indices (s,t) we have $y_{st} \neq 0$, then it must be that $\mu = \lambda_t - \lambda_{k+s}$. Once this μ is determined, then for all other indices (i,j) at which $\lambda_j - \lambda_{k+i} \neq \mu$, it must be that $y_{ij} = 0$. In this way we obtain a total of $k(n-k)$ many eigenvalues of the form

$$\mu = \lambda_j - \lambda_{k+i}, \quad i = 1, \ldots, n-k, \quad j = 1, \ldots k. \tag{4.25}$$

Note that all these eigenvalues are negative.

Likewise, define

$$Z := V_-^\top \widetilde{H} \in \mathbb{R}^{k\times k}.$$

Premultiply (4.22) by $V_-^\top$. We obtain

$$(Z + Z^\top)\Lambda_- = \mu Z. \tag{4.26}$$

Therefore, we need to satisfy

$$\begin{cases} (z_{ij} + z_{ji})\lambda_i = \mu z_{ij}, \\ (z_{ij} + z_{ji})\lambda_j = \mu z_{ji}, \end{cases}$$

simultaneously for all $i, j = 1, \ldots, k$. Observe that if for a fixed pair of indices (s,t), $s \neq t$, we have $z_{st} = 0$, then it must be $z_{ts} = 0$ also. That is, the off-diagonal elements z_{st} and z_{ts} must be simultaneously either zero or nonzero. Suppose that for a fixed pair of indices (s,t), $s \neq t$, we have nonzero z_{st} and z_{ts}. Then

$$\frac{\lambda_s}{\lambda_t} = \frac{z_{st}}{z_{ts}}.$$

It follows that

$$\mu = \left(\frac{z_{st} + z_{ts}}{z_{st}}\right)\lambda_s = \lambda_s + \lambda_t. \tag{4.27}$$

Once this μ is determined, then for all other indices (i,j) at which $\lambda_i + \lambda_j \neq \mu$, it must be that $z_{ij} = z_{ji} = 0$. Since z_{st} and z_{ts} are interchangeable, these eigenvalues should be counted twice, i.e., they are of multiplicity 2. There are a total of $k(k-1)$ many eigenvalues in this form, all of which are negative. Finally, if for a fixed s we have $z_{ss} \neq 0$, then

$$\mu = 2\lambda_s, \tag{4.28}$$

and $z_{ij} = 0$ for all other indices (i,j). There are k many such negative eigenvalues.

In all, we have found all $n+nk$ eigenvalues of the Jacobian matrix $\mathbb{J}(\mathbf{x}^*, U^*)$ as follows:

$$\begin{cases} \lambda_1, \dots, \lambda_k, -\lambda_{k+1}, \dots, -\lambda_n, & \\ \lambda_j - \lambda_{k+i}, & i = 1, \dots, n-k, \quad j = 1, \dots k, \\ \lambda_s + \lambda_t, & s, t = 1, \dots k, \quad s \neq t \text{ (double)}, \\ 2\lambda_s & s = 1, \dots k. \end{cases} \tag{4.29}$$

By the assumption, all these eigenvalues are real and negative. Therefore, the equilibrium point $(\mathbf{x}^*, U^*)$ is an attractor for the differential system (4.1). ∎

For the projector flow $P(t)$, recall that its evolution can be described either by the MRDE (3.15) or by the double bracket equation (3.18). While the MRDE does not generally preserve the projector condition (3.17), the double bracket formulation ensures that this property is maintained throughout the flow. In exact mathematics, therefore, we have the following convergence result.

COROLLARY 4.6. *Suppose that* $(\mathbf{x}^*, P^*)$ *is an equilibrium point of the continuous subspace dynamical system* (4.2)*, where* $\nabla f(\mathbf{x}^*) = 0$*, and* P^* *is an orthonormal basis for the eigenspace associated with the* k *leftmost eigenvalues of* $\nabla^2 f(\mathbf{x}^*)$*. Then the equilibrium point* $(\mathbf{x}^*, P^*)$ *for the projector flow* (4.2) *is an attractor.*

Proof. We have argued in Lemma 3.4 that the projector flow in the double bracket form (3.18) is a rewriting of the Oja flow (3.10) with the relationship that $P(t) = U(t)U(t)^\top$. The convergence of the system (4.1) implies the convergence of the differential system (4.2). In particular, $P^* = U^* U^{*\top}$ is the ultimate projector. ∎

Because the system (4.3) is a hybrid scheme, a distinct methodology is required to establish stability under more rigorous conditions. These conditions pertain to the perturbation of eigenvalues and, more critically, the eigenvectors of a Hermitian matrix. Foundational works in this area include the seminal texts [24, 41], which discuss, among other topics, Rellich's perturbation theorem concerning Hermitian matrices of holomorphic functions [49]. In the theorem that follows, we will accept the assumption as a given, without delving into the specific circumstances under which it is applicable.

THEOREM 4.7. *Regard the function* $Q(t)$ *defined in* (3.22) *as a map* $Q(\mathbf{x}(t))$*. Assume that eigenprojector* $\mathscr{P}(\mathbf{x}) = Q(\mathbf{x})Q(\mathbf{x})^\top$ *in* (4.3) *is differentiable in* $\mathbf{x}$*. Then* $\mathbf{x}^*$ *is asymptotically stable.*

Proof. Rewrite the dynamical system for $\mathbf{x}(t)$ in (4.3) as an autonomous system

$$\dot{\mathbf{x}} = -W(\mathbf{x})\nabla f(\mathbf{x}), \tag{4.30}$$

where

$$W(\mathbf{x}) = I_n - 2Q(\mathbf{x})Q(\mathbf{x})^\top.$$

The local stability of $\mathbf{x}^*$ depends on the eigenvalues of the Jacobian matrix of $-W(\mathbf{x})$ $\nabla f(\mathbf{x})$ at $\mathbf{x}^*$. We calculate the action of the Fréchet derivative at an arbitrary $\mathbf{h} \in \mathbb{R}^n$ as

$$\frac{\partial (W(\mathbf{x})\nabla f(\mathbf{x}))}{\partial \mathbf{x}}.\mathbf{h} = \left(\frac{\partial W(\mathbf{x})}{\partial \mathbf{x}}.\mathbf{h}\right)\nabla f(\mathbf{x}) + W(\mathbf{x})\nabla^2 f(\mathbf{x}).\mathbf{h},$$

where $\frac{\partial W}{\partial \mathbf{x}}$ is an order-3 tensor whose action at $\mathbf{h}$ produces a matrix. It follows that

$$\frac{\partial (-W(\mathbf{x})\nabla f(\mathbf{x}))}{\partial \mathbf{x}}|_{\mathbf{x}^*} = -W(\mathbf{x}^*)\nabla^2 f(\mathbf{x}^*).$$

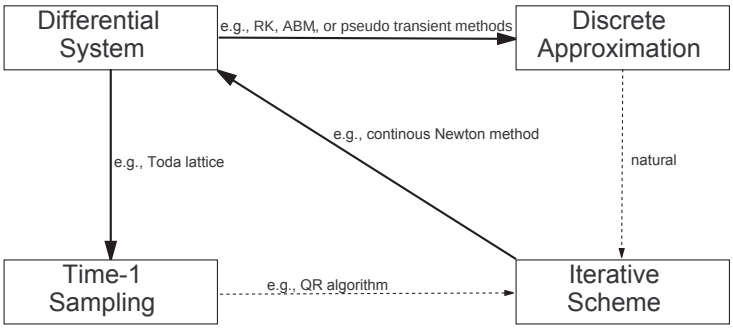

FIG. 2. *Possible interplay between continuous and discrete dynamical systems.*

Since we already know what $W(\mathbf{x}^*)$ is, we are in the same situation as that in Lemma 4.4. ∎

5. Numerical experiments. When the index k of a saddle point is known beforehand, we have introduced three dynamic systems designed to locate this saddle point. The theoretical exploration of their convergence is intriguing in its own right. In this section, we outline the practical aspects of implementing these methods numerically, discuss promising directions for future research, and present several empirical observations that illuminate their performance.

5.1. Numerical methods. In our extensive 86-page survey article [13], we have meticulously explored the intricate interplay between classical iterative methods and the dynamics of certain differential systems. Figure 2 provides a visual summary of the possible relationships between continuous and discrete dynamical systems. The uppermost pathway depicts the conventional approach: discretizing a differential system via sophisticated numerical integrators, thereby transforming the continuum of solutions onto a finite sequence of points. The leftward connection highlights instances where discrete models genuinely sample the solution trajectory of their corresponding differential equations at integer time steps, as exemplified by the correspondence between the QR algorithm and the Toda lattice [12, 47]. Dotted lines in the diagram suggest that iterative schemes may themselves be inspired by, or reconstructed from, underlying differential systems. While the transition from continuous to discrete is often considered natural, reflecting the spirit of most numerical ODE techniques, there exist unconventional discretization strategies, such as those in [26, 36, 46], which depart from standard ODE schemes and can yield innovative and more effective algorithms for special types of flows. Conversely, associating an iterative scheme with a differential system, such as the continuous power method (2.2) or the well-studied continuous Newton method, could be considerably richer than merely viewing the discrete method as an Euler step. Alternative mechanisms, including control, acceleration, optimization, and structure preservation, may also give rise to continuous dynamical systems, thereby transcending this basic interpretation and enriching the landscape of algorithmic design.

A central difficulty in this process of discretizing continuous systems is that it often compromises key mathematical features inherent to the continuous system, such as the asymptotic stability, the conservation of energy, or the preservation of geometric structure [19, 23, 33]. As a result, a raw discrete approximation may fail to faithfully reproduce the true behavior of the original model. Addressing this gap often requires

further refinement or modification of the numerical scheme, so that the computed solution more accurately reflects the dynamics of the continuous system. Contemporary numerical integrators typically utilize adaptive step sizes and variable-order methods to carefully track error propagation and maintain stability. Through these strategies, the resulting discrete solutions often closely mirror the dynamics of the underlying differential equations. Nevertheless, certain scenarios may arise in which this fidelity is not fully achieved.

To approximate the GAD originally described in the differential equation form (1.1), a variety of discrete-time schemes have been proposed in the literature. Among these, the iterative scheme proposed in [18] constructs an auxiliary minimization problem at each step, utilizing the minimal eigenmode of the Hessian. This approach achieves quadratic convergence near saddle points, but its inherently static nature means it does not reflect the temporal progression of the trajectory $\mathbf{x}(t)$. Moreover, its convergence guarantees are restricted to local neighborhoods, leaving the global behavior of the dynamics largely unexplored. On the other hand, the approach presented in [52] employs the Euler method to discretize their HiOSD system (3.1), originally introduced in [51]. While an error estimate for the trajectory approximation is provided, the Euler method is a rudimentary choice for integrating ODEs. The numerical analysis literature offers a wealth of more sophisticated integration schemes, which are known to deliver enhanced stability and accuracy. Employing such methods could substantially improve performance in this context. Algorithms based on the locally optimal block preconditioned conjugate gradient technique [6, 32] have also been explored. These algorithms require the computation of either exact or approximate eigenvectors of the Hessian. It has been observed that, at each iteration, the k eigenvectors corresponding to the smallest eigenvalues must be determined with high precision; otherwise, the algorithms may fail to converge.

In contrast, although our primary focus is on continuous-time dynamics rather than their discrete analogues, we must also emphasize that our systems are explicitly structured and ready for numerical integration, a process that inherently introduces discretization. In our numerical experiments, we employ the MATLAB ODE suite with exceptionally stringent error tolerances, setting both error controllers RelTol and AbsTol to 10^{-12}, thereby ensuring that the computed trajectories faithfully approximate their continuous counterparts. These advanced integrators employed are grounded on robust theoretical foundations and have proven reliability, allowing the numerical discrete solutions to reflect the behavior anticipated by the underlying mathematical theory for the continuous system. A systematic development of discrete analogues to our continuous methods remains an interesting direction for future research. Any ongoing progress in this direction should further enrich the field and deepen our understanding of saddle points.

5.2. Empirical results. For illustrative purposes, we have selected examples that are straightforward enough to allow precise identification of their saddle points and corresponding orientations. This prior knowledge serves to highlight the fascinating dynamics at play. The chosen problems exhibit diverse characteristics, which effectively demonstrate our points.

5.2.1. Surfaces with multiple index-1 saddle points and trajectory plots. We begin by examining several nontrivial surfaces embedded in $\mathbb{R}^3$, each featuring multiple saddle points of index $k = 1$. To illustrate the underlying dynamics, we present a collection of representative trajectories on these surfaces. Our observations reveal that convergence to the saddle points is inherently local, with trajectories approaching these points only from specific regions of the surface.

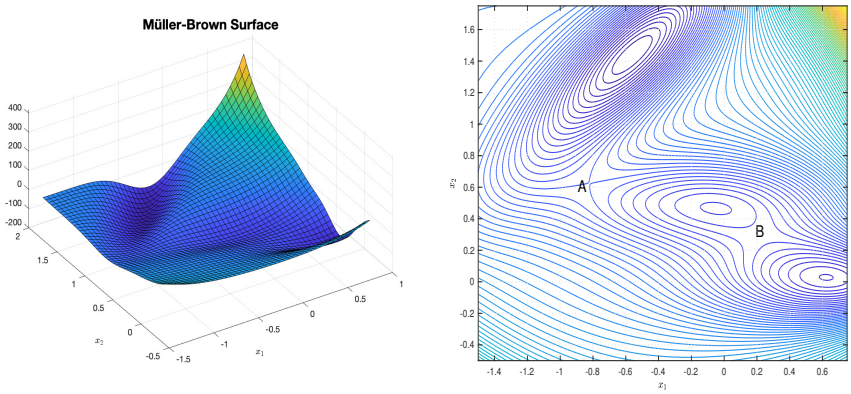

FIG. 3. *Müller–Brown potential energy surface and level curves.*

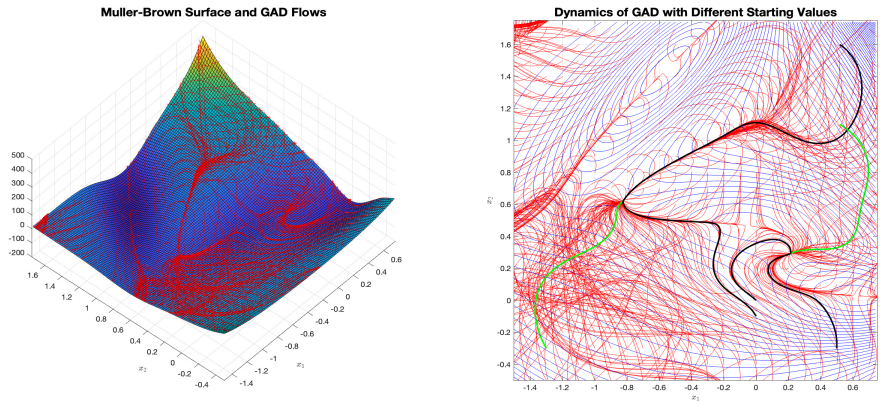

FIG. 4. *Dynamics of* (1.1) *applied to the Müller–Brown potential energy surface* (5.1).

Example 1. The Müller–Brown potential energy surface

$$
\begin{aligned}
f(x_1,x_2) = &-200e^{-(x_1-1)^2-10x_2^2} - 100e^{-x_1^2-10(x_2-0.5)^2} \\
&- 170e^{-6.5(x_1+0.5)^2+11(x_1+0.5)(x_2-1.5)-6.5(x_2-1.5)^2} \\
&+ 15e^{0.7(x_1+1)^2+0.6(x_1+1)(x_2-1)+0.7(x_2-1)^2}
\end{aligned} \tag{5.1}
$$

is a well-studied surface which has three local minima and two saddle points, as can be seen from the drawings in Figure 3. Both saddle points A and B are relatively shallow when comparing to their surroundings.

Starting with a grid of initial points for $\mathbf{x}(0)$ over the window $[-1.5, 0.75] \times [-0.5, 1.75]$ and a fixed $\mathbf{v}(0) = [1, 0]^\top$, we sketch some trajectories to demonstrate the ebb and flow of $\mathbf{x}(t)$ governed by (1.1) in Figure 4. Since the plot for $\mathbf{v}(t)$ is suppressed, the trajectories of $\mathbf{x}(t)$ may cross each other.

Naturally, altering the initial vector $\mathbf{v}(0)$ leads to different dynamical behaviors. We do not know of a strategy for selecting $\mathbf{v}(0)$ to enhance the convergence. In Figure 4, we observe that certain trajectories coalesce, suggesting underlying convergence in the flow. To clarify these dynamics, the right panel of Figure 4 highlights

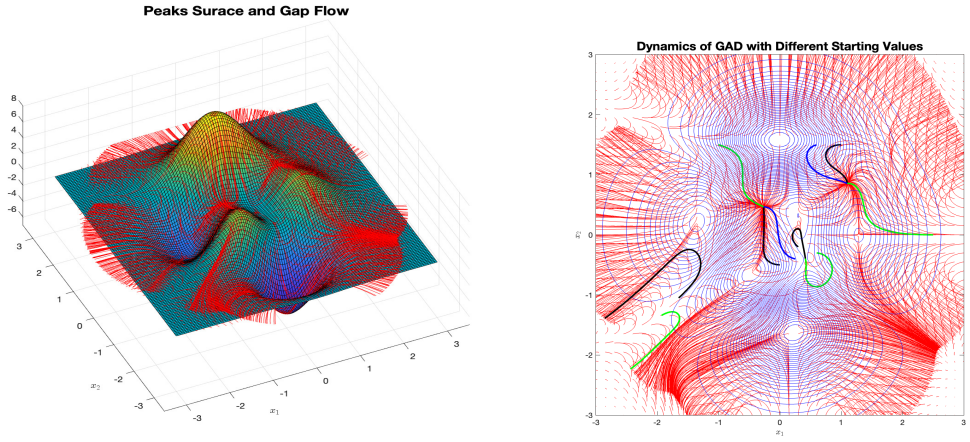

FIG. 5. *Landscape and dynamics of* (1.1) *applied to the peak function* (5.2).

several representative paths, each originating from distinct initial conditions and converging to saddle points. Meanwhile, the left panel offers a more vivid depiction of the "streams" formed by the GAD flows—some ascend and descend the landscape, while others trace the valleys. Notably, there exist trajectories that never reach the saddle points. The precise boundaries of the basins of attraction are challenging to delineate, as they depend not only on the initial position but also on the effect of $\mathbf{v}^*$. Nevertheless, both our theoretical analysis and these numerical experiments confirm that each pair $(\mathbf{x}^*, \mathbf{v}^*)$ is asymptotically stable. It is also evident from the trajectory plot that all three local minima act as repellers for the flow.

Example 2. The peak function used in MATLAB,

(5.2)
$$f(x_1,x_2) = 3(1-x_1)^2 e^{-x_1^2-(x_2+1)^2} - 10\left(\frac{x_1}{5} - x_1^3 - x_2^5\right) e^{-x_1^2-x_2^2} - \frac{1}{3} e^{-(x_1+1)^2-x_2^2},$$

has three maxima, three minima, and three saddle points. See the left drawing in Figure 5. Again, starting with a grid of points over the window $[-3,3] \times [-3,3]$ for $\mathbf{x}(0)$ and a fixed $\mathbf{v}(0) = [1,0]^\top$, we sketch some trajectories of $\mathbf{x}(t)$ governed by (1.1) in Figure 5. In the right panel of Figure 5, several illustrative trajectories are highlighted, including two (depicted in black and green) near the lower left corner that execute U-turns before ultimately diverging to infinity. This behavior underscores that, while the flow may at times ascend from a local minimum or descend from a local maximum, the GAD system does not universally guarantee that trajectories originating far from a saddle point will be drawn toward it. Although the flow $\mathbf{v}(t)$ is ensured to converge, for the position variable $\mathbf{x}(t)$, only local convergence to the saddle point can be expected in general.

5.2.2. Higher-index problems. In this subsection, we consider several problems characterized by higher index values. Additionally, we present an initial assessment of the computational performance of our three proposed dynamical systems.

Example 3. The 3-dimensional functional

(5.3)
$$w = f(x_1,x_2,x_3) = x_1^2 - x_2^2 - x_3^2 + x_1^3 - 3x_1x_2^2 - 3x_1x_3^2 - 3x_2x_3^2$$

has two saddle points $(0,0,0)$ and $(-\frac{2}{3},0,0)$ with the Hessians $\mathrm{diag}(2,-2,-2)$ and $\mathrm{diag}(-2,2,2)$, respectively. Thus, $(0,0,0)$ is an index-2 saddle point and $(-\frac{2}{3},0,0)$ is an index-1 saddle point, which makes this example an interesting test problem for the choice of k. For a given value w_0, the set

$$\{(x_1,x_2,x_3)|x_1^2-x_2^2-x_3^2+x_1^3-3x_1x_2^2-3x_1x_3^2-3x_2x_3^2=w_0\}$$

denotes a surface in $\mathbb{R}^3$. Since both saddle points reside in the (x_1,x_2)-plane, we report our empirical results in this plane only. The cross sections of different w-level surfaces with the (x_1,x_2)-plane are plotted in the top graph of Figure 6 where the two saddle points are marked as A and B, respectively. Starting with a grid of points in the (x_1,x_2)-plane with $x_3(0)=0$ together with a randomly generated $\mathbf{v}(0)\in\mathcal{S}(n,k)$, $k=1,2$, per grid point, we sketch some trajectories of $(x_1(t),x_2(t))$ in the bottom graphs of Figure 6, respectively. Note that $\mathbf{x}(t)$ should be a space curve in $\mathbb{R}^3$ in general. The effect of the correct choice of k is obvious. In the case $k=1$, we see that the saddle point B can be reached more frequently than A. Likewise, in the case $k=2$, the saddle point A is more reachable than B. Since $\mathbf{v}(0)$ is generate randomly, occasionally a wrong choice of k can still find the other saddle point, but the success is rare.

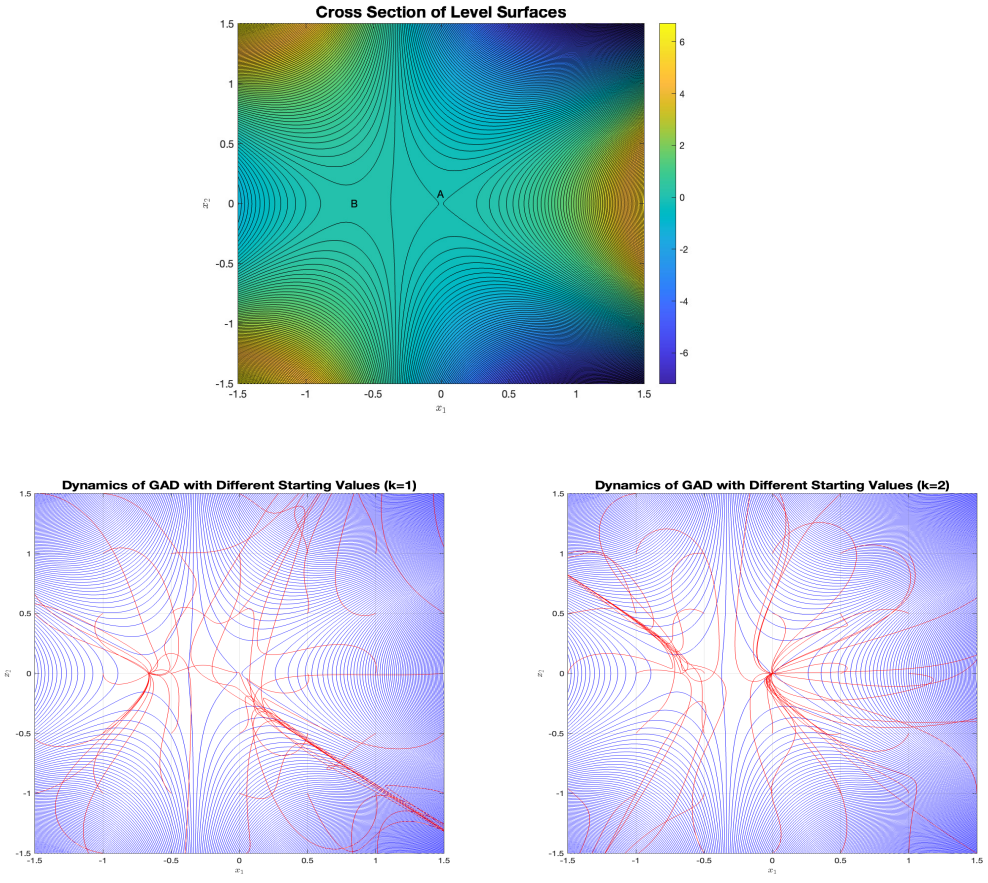

FIG. 6. *Landscape and dynamics of* (4.1) *with different* k *applied to the function* (5.3).

TABLE 1
Some saddle points for $B_5(\mathbf{x})$ and performance measurements.

Index k	5	4					3					2
ID	S_1	S_2	S_3	S_4	S_5	S_6	S_7	S_8	S_9	S_{10}	S_{11}	S_{12}
	1	-0.1351	1.0940	1.0125	0.8294	0.9806	-0.8725	-0.2132	-0.7749	-0.8906	-0.3453	-1.0814
	10	10.0153	5.5407	9.9985	10.0053	9.9564	4.4600	9.9847	9.9733	10.0001	9.9520	6.0377
saddle	1	1.1885	0.9092	1.0393	1.0800	0.9676	1.0131	1.1128	1.0368	1.0811	1.0848	0.9575
point	5	4.9863	5.0607	-0.3455	5.0030	10.4833	5.0990	5.0386	5.0444	-5.7751	9.7124	5.1237
	4	3.8584	4.1709	3.9871	1.7996	4.0266	4.0569	1.3024	10.7518	4.0080	3.9411	7.5968
	3	2.6429	4.3701	1.5290	2.0388	4.6195	3.6410	1.0733	3.1676	-2.5379	3.3152	3.5930
(4.1)	D	D	G	G	G	G	G	G	G	G	G	G
CPU	20.260	5.8576	5.7545	5.0930	3.3962	3.1055	2.8698	2.8705	3.3899	4.2244	3.3827	3.2002
(4.2)	D	D	G	G	G	G	G	G	G	G	G	G
CPU	3.6139	3.8315	3.4573	3.8078	4.6256	4.1719	3.5318	4.2147	3.9761	3.7425	3.5535	4.2656
(4.3)	D	D	G	G	G	G	G	G	$G^{\#}$	G	G	G
CPU	5.9611	3.5787	5.3785	5.3174	4.4294	5.1609	4.4975	4.7481	16.147	5.0212	3.6805	3.7405

Example 4. Consider the objective functions $B_k : \mathbb{R}^6 \to \mathbb{R}$ defined by [51]

(5.4)
$$B_k(\mathbf{x}) := B(\mathbf{x}) - \sum_{i=1}^{k} s_i \arctan^2(x_i - x_i^*) + \sum_{i=k+1}^{6} s_i \arctan^2(x_i - x_i^*), \quad k = 2, \dots 6,$$

where

$$B(\mathbf{x}) := \sum_{i=1}^{6} (x_3 e^{-t_i x_1} - x_4 e^{-t_i x_2} + x_6 e^{-t_i x_5} - y_i)^2,$$

with

$$\begin{cases} t_i = \dfrac{i}{10}, \\ y_i = e^{-t_i} - 5e^{-10t_i} + 3e^{-4t_i}, \end{cases} \qquad i = 1, \dots 6,$$
$$\mathbf{s} = [4, 8, 16, 8, 4, 2]^\top,$$
$$\mathbf{x}^* = [1, 10, 1, 5, 4, 3]^\top.$$

It can be verified that for $k = 2, \dots, 6$, the same $\mathbf{x}^*$ is a saddle point of index k for $B_k(\mathbf{x})$. Indeed for $B_5(\mathbf{x})$ alone, we have found many other saddle points. A partial list of these points together with their indices is displayed up to four digits in Table 1.

For demonstration, we report the performance test of our three dynamical systems on the problem $B_5(\mathbf{x})$ only at the bottom of Table 1. To maintain uniformity across different domains of attraction for each point, we initiate all three dynamical systems from a common starting point selected from the same vicinity in the sense of

$$\mathbf{x}_0 = S_j + \mathcal{N}(0, \sigma^2).$$

Here, $\mathcal{N}(0, \sigma^2)$ represents a random vector drawn from a Gaussian distribution with a standard deviation of $\sigma = 0.1$. The term S_j refers to one of the 12 saddle points listed in Table 1. The integration process is terminated when one of the stopping criteria is satisfied, identified as either "G" or "D," indicating the gradient norm or the distance to the targeted S_j is less than 10^{-10}. We thus record the CPU time in the units of 10^{-2} seconds. For the scheme (4.3), we fix $c = 30$ in (3.23) to ensure $C(\mathbf{x}(t))$ remains positive definite. The QR updates are limited to 10 iterations per time step which is

TABLE 2
Convergence of continuous subspace flow with distinct eigenspace dimensions.

	True index →	5	4				3						2
		S_1	S_2	S_3	S_4	S_5	S_6	S_7	S_8	S_9	S_{10}	S_{11}	S_{12}
dimension p	$p=2$	X	X	X	G*	X	X	G*	G*	X	G*	X	S_{12}
	$p=3$	S_9	S_8	S_7	X	S_8	G*	S_7	S_8	S_9	S_{10}	S_{11}	G*
	$p=4$	S_5	S_2	S_3	S_4	S_5	S_6	S_2	G*	S_2	S_2	S_2	S_2
	$p=5$	S_1	X	S_1	S_1	X	S_1	X	X	X	S_4	S_1	X

determined by the `ode15s` routine in the MATLAB ODE suite. We observe that in most instances where the "G" event occurs, the calculated distance falls between 10^{-9} and 10^{-10}, with one notable exception. This exception, marked as G$^{\#}$, involves the saddle point S_9 whose distance to the limit point is approximately 8.2695, while the gradient norm is less than 10^{-10}. This indicates that convergence to an alternative saddle point, even when starting near a specific one, is possible and warrants further investigation. What has happened is that the limitation of updates to 10 iterations may have hindered the proper identification of the dominant eigenspace. Interestingly, increasing the iterations to 11 allows the original saddle point to be reached. Similarly, reducing the standard deviation of the perturbation achieves the same result. These observations support our theory of local convergence.

In practice, the exact locations of saddle points are not known beforehand. In particular, we have no a priori knowledge of their indices. In our subsequent experiment, we maintain the initial setup to explore how choosing dimensions $p = 2, 3, 4, 5$ for the dominant eigenspaces influences convergence. Given that convergence is not ensured, we extend the ODE solver's integration interval to a maximum of 10^3. If no stopping criterion is met within this timeframe, an "X" is recorded in Table 2 to indicate nonconvergence. Beyond the convergence results already presented in Table 1, we observe that initiating near Sj with index k but selecting an incorrect dimension p may lead to nonconvergence. If convergence occurs, the flow identifies a limit point with an index exactly matching p. Additionally, the notation G* in Table 2 signifies the discovery of a "new" saddle point with index p while satisfying the gradient condition.

5.2.3. A saddle point problem arising from PDE. Saddle points arise naturally in the analysis of nonlinear partial differential equations, where the equation is cast as the necessary condition of a variational problem. When the solution does not correspond to an extremum, it is characterized as a saddle point of the associated functional, and often there are multiple high-index saddle points.

Example 5. Consider the functional

$$E(u) := \int_\Omega \left(\frac{1}{2} \|\nabla u(\mathbf{x})\|^2 - \frac{1}{4} u^4(\mathbf{x}) \right) d\mathbf{x} \tag{5.5}$$

for $u \in H_0^1(\Omega)$, where $\Omega \subset \mathbb{R}^2$ is a bounded Lipschitz domain. For simplicity, we shall employ the natural L^2 inner product only over $H_0^1(\Omega)$. Regarding the gradient of $E(u)$ as a linear functional acting on $H_0^1(\Omega)$, we may write

$$\nabla E(u) = -u^3 - \Delta u, \tag{5.6}$$

where Δ denotes the Laplacian operator. The Euler–Lagrange equation

$$-\Delta u = u^3, \quad u|_{\partial\Omega} = 0 \tag{5.7}$$

specifies the necessary condition for a critical point. This special Lane–Emden equation is known to have multiple solutions which do not have the analytic form. Any solution to (5.7) which is not an extreme point is a saddle point of the objective functional $E(u)$. The Fréchet derivative $\nabla^2 E(u)$ of ∇E is a linear transformation whose acting on $\eta \in H_0^1(\Omega)$ assumes the form

$$\nabla^2 E(u).\eta = -3u^2\eta - \Delta\eta. \tag{5.8}$$

The eigenvalues of $\nabla^2 E(u)$ should depend on u. Solving this eigenvalue problem, even with a specified $u(\mathbf{x})$, presents considerable challenges.

Instead, by applying a proper finite difference or finite element scheme, we may assume that the function u over Ω has been discretized to a column vector $\mathbf{u}$ and the Laplacian operator $-\Delta$ has been approximated by a fixed matrix L. We thus may regard

$$\nabla E(\mathbf{u}) = -\mathbf{u}.^3 + L\mathbf{u}, \tag{5.9}$$

$$\nabla^2 E(\mathbf{u}) = -3\,\mathrm{diag}(\mathbf{u}.^2) + L \tag{5.10}$$

as the approximate gradient and Hessian to (5.7) and (5.8), respectively, where $\mathbf{u}.^3$ denotes the elementwise cubic of $\mathbf{u}$ and $\mathrm{diag}(\mathbf{u}.^2)$ denotes the diagonal matrix with the elementwise quadratic $\mathbf{u}.^2$ along the diagonal. While these approximations are subject to truncation errors, we now possess all the necessary components to construct the GAD. The finer the discretization mesh adopted, the better the eigenvalues are approximated. What is important is that the index k of $\nabla^2 E(u)$ aligns with that of $\nabla^2 E(\mathbf{u})$, provided that the mesh size is sufficiently small [48].

To illustrate, we examine the ODE variant of the preceding analysis. The eigenvalue problem is given by

$$-\eta'' - 3u^2\eta = \lambda\eta, \quad \eta(0) = \eta(1) = 0, \tag{5.11}$$

which represents a special "regular" Sturm–Liouville problem, where $u(x)$ satisfies

$$-u'' = u^3, \quad u(0) = u(1) = 0. \tag{5.12}$$

According to the Sturm–Liouville theory [39], the eigenvalues are real and can be enumerated as

$$\lambda_1(u) < \lambda_2(u) < \cdots < \lambda_n(u) < \cdots \to \infty,$$

and corresponding to each eigenvalue λ_n is a unique (up to constant multiple) eigenfunction $\eta_n(x)$ with exactly $n-1$ zeros in the interval $[0,1]$. Moreover, the set $\{\eta_n\}$ forms an orthonormal basis under the inner product in the Hilbert space $L^2[0,1]$. With this in mind, the L matrix by the 3-point central difference formula is the symmetric Toeplitz matrix with first row given by $[2,-1,0,\ldots]$. With $h = \frac{1}{500}$, for instance, we identify saddle points with notably high indices. Figure 7 illustrates three such solutions, with indices of 5, 7, and 10, respectively.

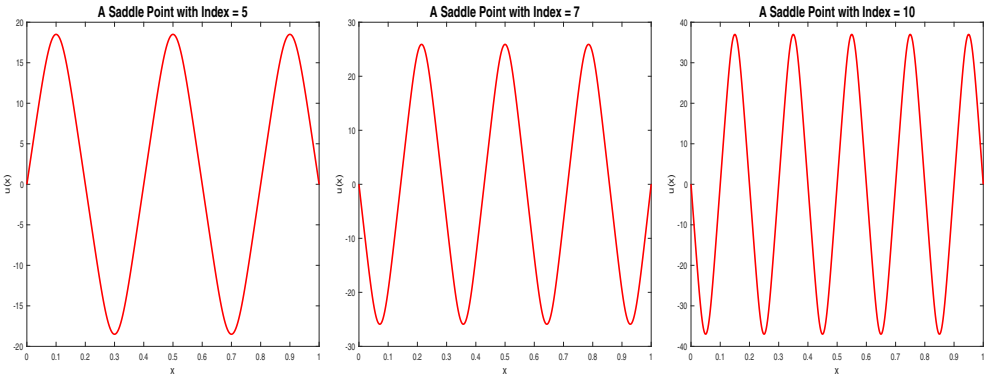

FIG. 7. *Saddle points to $E(u)$ (ODE variant) with high indices.*

6. Conclusion. We reinterpret the gentlest ascent dynamics for index-1 saddle point problems as the Householder reflection of the negative gradient, guided by a vector from the continuous power method, Such an adaptation strategy adheres to a specific criterion. Leveraging this geometric insight, the paper introduces three generalized dynamical systems for high-index saddle point problems, focusing on estimating directions to construct the generalized reflector. The first method employs the Oja flow for eigenspaces, with the continuous power method as a special case. The second method works with the projector operator on the Grassmann manifold, which is shown to be equivalent to a double bracket flow. The third method is a hybrid approach using conventional subspace iteration with QR factorization for normalization. The equilibrium points of these systems are classified, and convergence analysis is provided for each. Numerical experiments validate the theoretical framework.

REFERENCES

[1] P.-A. Absil, R. Sepulchre, and R. Mahony, *Continuous-time subspace flows related to the symmetric eigenproblem*, Pacific J. Optim., 4 (2008), pp. 179–194.

[2] W. Arnold and A. Laub, *Generalized eigenproblem algorithms and software for algebraic Riccati equations*, Proc. IEEE, 72 (1984), pp. 1746–1754.

[3] J. M. Bello-Rivas, A. Georgiou, H. Vandecasteele, and I. G. Kevrekidis, *Gentlest ascent dynamics on manifolds defined by adaptively sampled point-clouds*, J. Phys. Chem. B, 127 (2023), pp. 5178–5189, https://doi.org/10.1021/acs.jpcb.3c00897.

[4] A. Bloch, R. Brockett, and T. Ratiu, *On the geometry of saddle point algorithms*, in Proceedings of the 31st IEEE Conference on Decision and Control, Vol. 2, 1992, pp. 1482–1487.

[5] A. M. Bloch and A. Iserles, *On the optimality of double-bracket flows*, Int. J. Math. Math. Sci., 2004 (2004), pp. 3301–3319, https://doi.org/10.1155/S0161171204406462.

[6] J. M. Bofill, J. Ribas-Ariño, R. Valero, G. Albareda, I. de P. R. Moreira, and W. Quapp, *Interplay between the gentlest ascent dynamics method and conjugate directions to locate transition states*, J. Chem. Theory Comput., 15 (2019), pp. 5426–5439, https://doi.org/10.1021/acs.jctc.8b01061.

[7] S. Boyd and L. Vandenberghe, *Convex Optimization*, Cambridge University Press, Cambridge, UK, 2004.

[8] R. W. Brockett, *Dynamical systems that sort lists, diagonalize matrices, and solve linear programming problems*, Linear Algebra Appl., 146 (1991), pp. 79–91, https://doi.org/10.1016/0024-3795(91)90021-N.

[9] M. Cameron, R. V. Kohn, and E. Vanden-Eijnden, *The string method as a dynamical system*, J. Nonlinear Sci., 21 (2011), pp. 193–230, https://doi.org/10.1007/s00332-010-9081-y.
[10] F. Casas, *Numerical integration methods for the double-bracket flow*, J. Comput. Appl. Math., 166 (2004), pp. 477–495, https://doi.org/10.1016/j.cam.2003.08.046.
[11] C. J. Cerjan and W. H. Miller, *On finding transition states*, J. Chem. Phys., 75 (1981), pp. 2800–2806, https://doi.org/10.1063/1.442352.
[12] M. T. Chu, *On the continuous realization of iterative processes*, SIAM Rev., 30 (1988), pp. 375–387, https://doi.org/10.1137/1030090.
[13] M. T. Chu, *Linear algebra algorithms as dynamical systems*, Acta Numer., 17 (2008), pp. 1–86, https://doi.org/10.1017/S0962492906340019.
[14] M. T. Chu and K. R. Driessel, *The projected gradient method for least squares matrix approximations with spectral constraints*, SIAM J. Numer. Anal., 27 (1990), pp. 1050–1060, https://doi.org/10.1137/0727062.
[15] G. Crippen and H. Scheraga, *Minimization of polypeptide energy:* XI. *The method of gentlest ascent*, Arch. Biochem. Biophys., 144 (1971), pp. 462–466, https://doi.org/10.1016/0003-9861(71)90349-3.
[16] W. E, W. Ren, and E. Vanden-Eijnden, *Simplified and improved string method for computing the minimum energy paths in barrier-crossing events*, J. Chem. Phys., 126 (2007), https://doi.org/10.1063/1.2720838.
[17] W. E and X. Zhou, *The gentlest ascent dynamics*, Nonlinearity, 24 (2011), pp. 1831–1842, https://doi.org/10.1088/0951-7715/24/6/008.
[18] W. Gao, J. Leng, and X. Zhou, *An iterative minimization formulation for saddle point search*, SIAM J. Numer. Anal., 53 (2015), pp. 1786–1805, https://doi.org/10.1137/130930339.
[19] E. Hairer, C. Lubich, and G. Wanner, *Geometric Numerical Integration*, 2nd ed., Springer Ser. Comput. Math. 31, Springer-Verlag, Berlin, 2006.
[20] P. Hänggi, P. Talkner, and M. Borkovec, *Reaction-rate theory: Fifty years after Kramers*, Rev. Modern Phys., 62 (1990), pp. 251–341, https://doi.org/10.1103/RevModPhys.62.251.
[21] G. Henkelman, G. Jóhannesson, and H. Jónsson, *Methods for finding saddle points and minimum energy paths*, in Progress on Theoretical Chemistry and Physics, S. Schwartz, ed., Kluwer, Dordrecht, 2006, pp. 269–302.
[22] A. Iserles, *On the discretization of double-bracket flows*, Found. Comput. Math., 2 (2002), pp. 305–329, https://doi.org/10.1007/s102080010024.
[23] A. Iserles, H. Z. Munthe-Kaas, S. P. Nørsett, and A. Zanna, *Lie-group methods*, Acta Numer., 9 (2000), pp. 215–365.
[24] T. Kato, *Perturbation Theory for Linear Operators*, Classics in Math. 132, Springer-Verlag, Berlin, 1995.
[25] V. N. Katsikis, S. D. Mourtas, P. S. Stanimirović, and Y. Zhang, *Continuous-time varying complex QR decomposition via zeroing neural dynamics*, Neural Process. Lett., 53 (2021), pp. 3573–3590, https://doi.org/10.1007/s11063-021-10566-y.
[26] C. T. Kelley, L.-Z. Liao, L. Qi, M. T. Chu, J. P. Reese, and C. Winton, *Projected pseudotransient continuation*, SIAM J. Numer. Anal., 46 (2008), pp. 3071–3083, https://doi.org/10.1137/07069866X.
[27] Y.-C. Kuo, H.-E. Lin, and S.-F. Shieh, *Asymptotic analysis of generalized orthogonal flows*, Linear Algebra Appl., 717 (2025), pp. 1–25, https://doi.org/10.1016/j.laa.2025.03.018.
[28] V. Kučera, *A review of the matrix Riccati equation*, Kybernetika (Prague), 9 (1973), pp. 42–61.
[29] A. Laub, *A Schur method for solving algebraic Riccati equations*, IEEE Trans. Automat. Control, 24 (1979), pp. 913–921, https://doi.org/10.1109/TAC.1979.1102178.
[30] J. J. Levin, *On the matrix Riccati equation*, Proc. Amer. Math. Soc., 10 (1959), pp. 519–524, https://doi.org/10.1090/S0002-9939-1959-0108628-X.
[31] W. Liu, Z. Xie, and Y. Yuan, *A constrained gentlest ascent dynamics and its applications to finding excited states of Bose-Einstein condensates*, J. Comput. Phys., 473 (2023), 111719, https://doi.org/10.1016/j.jcp.2022.111719.
[32] Y. Luo, X. Zheng, X. Cheng, and L. Zhang, *Convergence analysis of discrete high-index saddle dynamics*, SIAM J. Numer. Anal., 60 (2022), pp. 2731–2750, https://doi.org/10.1137/22M1487965.
[33] R. I. McLachlan, G. R. W. Quispel, and N. Robidoux, *Geometric integration using discrete gradients*, Philos. Trans. A, 357 (1999), pp. 1021–1045, https://doi.org/10.1098/rsta.1999.0363.
[34] J. Milnor, *Morse Theory*, Ann. Math. Stud. 51, Princeton University Press, Princeton, NJ, 1963.

[35] J. Nocedal and S. J. Wright, *Numerical Optimization*, 2nd ed., Springer Ser. Oper. Res. Financ. Eng., Springer, New York, 2006.
[36] R. A. Norton and G. R. W. Quispel, *Discrete gradient methods for preserving a first integral of an ordinary differential equation*, Discrete Contin. Dyn. Syst., 34 (2014), pp. 1147–1170, https://doi.org/10.3934/dcds.2014.34.1147.
[37] E. Oja, *Neural networks, principal components, and subspaces*, Int. J. Neural Syst., 01 (1989), pp. 61–68, https://doi.org/10.1142/S0129065789000475.
[38] A. Pedersen, S. F. Hafstein, and H. Jónsson, *Efficient sampling of saddle points with the minimum-mode following method*, SIAM J. Sci. Comput., 33 (2011), pp. 633–652, https://doi.org/10.1137/100792743.
[39] J. D. Pryce, *Numerical Solution of Sturm-Liouville Problems*, Monographs on Numerical Analysis, Oxford University Press, New York, 1993.
[40] W. Quapp and J. M. Bofill, *Locating saddle points of any index on potential energy surfaces by the generalized gentlest ascent dynamics*, Theoret. Chem. Acc., 133 (2014), pp. 1–14, https://doi.org/10.1007/s00214-014-1510-9.
[41] F. Rellich, *Perturbation Theory of Eigenvalue Problems*, Gordon and Breach, New York, 1969.
[42] W.-Q. Ren, *Higher order string method for finding minimum energy paths*, Commun. Math. Sci., 1 (2003), pp. 377–384, https://doi.org/10.4310/CMS.2003.v1.n2.a10.
[43] Y. Saad, *Analysis of subspace iteration for eigenvalue problems with evolving matrices*, SIAM J. Matrix Anal. Appl., 37 (2016), pp. 103–122, https://doi.org/10.1137/141002037.
[44] A. Samanta and W. E, *Optimization-based string method for finding minimum energy path*, Commun. Comput. Phys., 14 (2013), pp. 265–275, https://doi.org/10.4208/cicp.220212.030812a.
[45] T.-Y. Tam, *Gradient flows and double bracket equations*, Differential Geom. Appl., 20 (2004), pp. 209–224, https://doi.org/10.1016/j.difgeo.2003.10.008.
[46] S. Tsujimoto, Y. Nakamura, and M. Iwasaki, *The discrete Lotka-Volterra system computes singular values*, Inverse Problems, 17 (2001), pp. 53–58, https://doi.org/10.1088/0266-5611/17/1/305.
[47] D. S. Watkins, *Understanding the QR algorithm*, SIAM Rev., 24 (1982), pp. 427–440, https://doi.org/10.1137/1024100.
[48] H. F. Weinberger, *Upper and lower bounds for eigenvalues by finite difference methods*, Comm. Pure Appl. Math., 9 (1956), pp. 613–623, https://doi.org/10.1002/cpa.3160090329.
[49] H. K. Wimmer, *Rellich's perturbation theorem on Hermitian matrices of holomorphic functions*, J. Math. Anal. Appl., 114 (1986), pp. 52–54, https://doi.org/10.1016/0022-247X(86)90064-8.
[50] W.-Y. Yan, U. Helmke, and J. Moore, *Global analysis of Oja's flow for neural networks*, IEEE Trans. Neural Networks, 5 (1994), pp. 674–683, https://doi.org/10.1109/72.317720.
[51] J. Yin, L. Zhang, and P. Zhang, *High-index optimization-based shrinking dimer method for finding high-index saddle points*, SIAM J. Sci. Comput., 41 (2019), pp. A3576–A3595, https://doi.org/10.1137/19M1253356.
[52] L. Zhang, P. Zhang, and X. Zheng, *Error estimates for Euler discretization of high-index saddle dynamics*, SIAM J. Numer. Anal., 60 (2022), pp. 2925–2944, https://doi.org/10.1137/21M1458314.