SIAM J. NUMER. ANAL.
Vol. 53, No. 4, pp. 1786–1805




# AN ITERATIVE MINIMIZATION FORMULATION FOR SADDLE POINT SEARCH*

WEIGUO GAO†, JING LENG‡, AND XIANG ZHOU§

**Abstract.** This paper proposes and analyzes an iterative minimization formulation for searching index-1 saddle points of an energy function. We give a general and rigorous description of eigenvector-following methodology in this iterative scheme by considering an auxiliary optimization problem at each iteration in which the new objective function is locally defined near the current guess. We prove that this scheme has a quadratic local convergence rate in terms of number of iterations, in comparison to the linear rate of the gentlest ascent dynamics [W. E and X. Zhou, *Nonlinearity*, 24 (2011), pp. 1831–1842] and many other existing methods. We also propose the generalization of the new methodology for saddle points of higher index and for constrained energy functions on the manifold. Preliminary numerical results on the nature of this iterative minimization formulation are presented.

**Key words.** saddle point, energy landscape, eigenvector-following, gentlest ascent dynamics, iterative minimization

**AMS subject classifications.** Primary, 65K05; Secondary, 82B05

**DOI.** 10.1137/130930339

## 1. Introduction.

For some time considerable attention has been given to numerical methods of searching local minima of a continuous nonlinear function. The widespread availability of the efficient optimization algorithms for large scale problems has greatly assisted the numerical studies of theoretical physics, chemistry, and biology. In computational chemistry, for example, it is of great interest to look for metastable states of molecular configurations, which correspond to local minima of an energy function. Normally, the traditional optimization procedures are very successful at locating a nearby metastable state. However, of more interest are the transition states in these molecular systems, which are the saddle points of the energy function. When it comes to the location of the transition states, there is much room for improvement in the minimization approach.

Transition states are characterized as stationary points having one, and only one, negative Hessian eigenvalues (e.g., see [25]). Saddle points of this type are usually referred to as index-1 saddle points. There have already been various advanced algorithms during the past decades that have proved to be efficient in searching saddles for many practical problems in chemistry and material sciences. The contributions

*Received by the editors July 23, 2013; accepted for publication (in revised form) May 7, 2015; published electronically July 16, 2015.
http://www.siam.org/journals/sinum/53-4/93033.html

†School of Mathematical Sciences, Fudan University, Shanghai, 200433 China, and MOE Key Laboratory of Computational Physical Sciences, Fudan University, Shanghai 200433, China (wggao@fudan.edu.cn). The research of this author was supported by the National Natural Science Foundation of China under grant 91330202, Shanghai Science and Technology Development Funds 13dz2260200 and 13511504300, and Special Funds for Major State Basic Research Projects of China (2015CB858560003).

‡School of Mathematical Sciences, Fudan University, Shanghai 200433, China (12110180015@fudan.edu.cn).

§Corresponding author. Department of Mathematics, City University of Hong Kong, Kowloon, Hong Kong SAR (xiang.zhou@cityu.edu.hk). The research of this author was supported by CityU Start-Up Grant (7200301) and grants from the Research Grants Council of the Hong Kong Special Administrative Region, China (Project No. CityU 11304314, 109113).

include, but are not limited to, the following list: the activation-relaxation techniques [20], the dimer method [13], the nudged elastic band method [15], and the string method [9, 10, 23]. Interested readers can also refer to [12]. The first two methods in the list are examples of "single-state" (or "surface walker") algorithms, and the last two are, roughly speaking, examples of "chain-of-state" algorithms. The single-state algorithms mainly adopt the "eigenvector-following" methodology [7, 6, 25]—the system is moved uphill along the eigenvector ("min-mode") corresponding to the smallest eigenvalue of the Hessian matrix. Therefore, these methods drive the system away from the local minimum and push it to some index-1 saddle point if the convergence is achieved. Numerous applications to practical problems have shown that these eigenvector-following–type (or min-mode) methods generally have a much larger attraction domain for convergence to index-1 saddle points than the traditional Newton-type root-finding methods. In addition, the specificity of selecting index-1 saddles renders these methods more favorable than the root-finding methods. One more benefit of using eigenvector-following ideas over the Newton root-finding method is that the explicit information of Hessian matrix is usually not required in numerical implementation. Lastly, when the Hessian is quite close to a singular matrix, the Newton-type method will have difficulty, but the eigenvector-following methodology which needs only the minimal mode of the Hessian does not suffer from this singularity; see details of these methods in section 3.1.

Recently, there has been increasing mathematical interest in writing the eigenvector-following methodology in the form of a dynamical system. For instance, one of the authors of the current paper has proposed the gentlest ascent dynamics (GAD) [11, 24], which is a coupled dynamical system of both a position variable and a direction variable. A different but quite similar dynamical system is also pursued in [27] to implement the finite differencing by introducing one more dimer length variable.

In GAD, the dynamics flow is defined on the product space of the position in the configuration space and the direction in its tangent space. The position variable describes the escape trajectory from the basins of attraction of the local minima. The direction variable in GAD simultaneously evolves to try to follow the min-mode of the Hessian matrix, although it does not have to be exactly the min-mode at any time. It is proved that the stable equilibrium points of this dynamical system are the index-1 saddle points of the energy function, while the local minima of the energy function are turned into the index-1 saddle points of GAD. This interesting property invites one to attempt to think of GAD as a counterpart of the very basic steepest descent dynamics (SDD), which converges to a local minimum as time goes to infinity. GAD and SDD are both the simplest flow based on the gradient of the energy function in the configuration space, and the convergence rates are both linear.

SDD is closely related to the steepest descent method, the simplest gradient method for unconstrained optimization, which can be traced back to Cauchy [5]. The analogy between SDD and GAD is a tentative attempt to compare the framework of various optimization algorithms and that of the saddle search algorithms. It is well known that the steepest descent method is ineffective for unconstrained optimization because of its slow convergence rate. Historically, many better alternative optimization techniques have been developed to achieve a superlinear or quadratic convergence rate, for instance, the Newton method, L-BFGS, the nonlinear conjugate gradient method, and so on [21]. We are interested in asking what could be the possible analogues of these advanced optimization methods in the context of saddle search problems and how to improve the linear convergence of the GAD as well as

other popular saddle search algorithms.

Our motivation is thus to address the above questions and includes the following twofold tasks. First, we want to present a new mathematical framework with connection to some optimization problem, rather than in the form of a dynamical system, with the hope that the GAD is a natural "gradient flow" of the associated optimization problem. Second, the new formulation should be able to provide a superlinear or quadratic convergence rate and carry more flexibility in designing faster algorithms. This paper focuses on the first goal in a theoretical aspect and offers a partial discussion of the second goal with preliminary numerical experiments. The full discussion of developing faster numerical methods and applying them to real problems will be presented in a separate paper.

The formulation we propose in this note is an iterative minimization scheme. At each iteration, a new objective function is constructed based on the given energy function by using the information of the current values of the position and the minimal mode of the Hessian. Then a *local* minimizer of this objective function is assigned to the new value of the position at the next iteration. This iterative scheme can be completely described by a continuously differentiable mapping. We discover that the index-1 saddle point is a fixed point of this mapping, and the Jacobian matrix of this mapping vanishes at the saddle point, suggesting that the convergence rate of the iterative minimization scheme is quadratic.

We note that a few variant techniques have been proposed in an effort to improve the efficiency of the single-state–type algorithms for saddle search, such as those in [14, 16, 4, 17]. However, all of these methods either improve the rotation step of solving the eigenvector or improve the translation step of moving the position restricted in the dimension along the obtained direction. The resulting overall effect on the accuracy of these methods is of only the linear convergence rate in the full configuration space.

The rest of this paper is organized as follows. First we briefly review the gentlest ascent dynamics in section 2. In section 3, we formulate our iterative minimization scheme for index-1 saddles and analyze the convergence rate. We also discuss the situation with constraints for saddle points in section 4. The generalization for index-$m$ ($m > 1$) saddles is discussed in section 5. Several numerical examples are presented in section 6 to illustrate our theory. Section 7 offers concluding remarks.

**2. Review of gentlest ascent dynamics (GAD).** The GAD is a mathematical model in the form of a dynamical system to describe the escape of the basin of attractions (in the gentlest way) and the convergence to a saddle. Given a smooth energy function $V$ on the configuration space, say $\mathbb{R}^d$, the GAD is the following dynamical system defined on the phase space $\mathbb{R}^d \times \mathbb{R}^d$:

$$\begin{cases} \dot{x} = -\nabla V(x) + 2\dfrac{\langle \nabla V(x), v\rangle}{\langle v, v\rangle} v, & \text{(2.1a)} \\ \gamma \dot{v} = -\nabla^2 V(x) v + \dfrac{\langle v, \nabla^2 V(x) v\rangle}{\langle v, v\rangle} v. & \text{(2.1b)} \end{cases}$$

Here $\langle \cdot, \cdot \rangle$ is the dot product in the Euclidean space $\mathbb{R}^d$ (there is no difficulty in generalizing to any Hilbert space), and the relaxation constant $\gamma$ can be any positive real number. The second equation (2.1b) attempts to find the direction that corresponds to the smallest eigenvalue of the Hessian matrix $\nabla^2 V(x)$. The second term in (2.1b) imposes the normalization condition that $\|v\| = \sqrt{\langle v, v\rangle} = 1$. The last term in the first equation (2.1a) reverses the component of the gradient force in the direction $v$ to drive the system uphill in the direction of $v$.

It is shown in [11] that the saddle point of the original function $V$ is the stable equilibrium point of the GAD. For the reader's convenience, we recall this result in the following proposition.

PROPOSITION 2.1. *Assume that the energy function $V$ is $\mathcal{C}^4(\mathbb{R}^d;\mathbb{R})$.*

(a) *If $(x_*, v_*)$ is an equilibrium point of the GAD (2.1) and $\|v_*\| = 1$, then $v_*$ is an eigenvector of $\nabla^2 V(x_*)$ corresponding to some eigenvalue $\lambda_*$, and $x_*$ is a stationary point of the steepest descent dynamics of $V$, i.e., $\nabla V(x_*) = 0$.*

(b) *Suppose that $x_s$ is a stationary point of $V$, i.e., $\nabla V(x_s) = 0$. Let $v_1, v_2, \ldots, v_d$ be the normalized eigenvectors of the Hessian $\nabla^2 V(x_s)$, and let the associated eigenvalues be $\lambda_1, \lambda_2, \ldots, \lambda_d$, respectively. Then for all $i = 1, \ldots, d$, $(x_s, v_i)$ is an equilibrium point of the GAD (2.1). Furthermore, among these $d$ equilibrium points, there exists one pair $(x_s, v_{i'})$ which is linearly stable for the GAD (2.1) if and only if $x_s$ is an index-1 saddle point of the function $V$ or, equivalently, the eigenvalue $\lambda_{i'}$ corresponding to $v_{i'}$ is the only negative eigenvalue of $\nabla^2 V(x_s)$.*

For notational convenience, we denote the Hessian as $H(x) \triangleq \nabla^2 V(x)$. When the GAD converges, the limit of $v$ corresponds to the eigenvector of the Hessian $H(x)$ at the saddle point for the smallest eigenvalue. Actually, for any frozen $x$ in (2.1b), the steady state of the solution $v(t)$ solves the minimization problem for the Rayleigh quotient,

$$\min_{\|u\|=1} u^{\mathsf{T}} H(x) u, \tag{2.2}$$

and (2.1b) is just a steepest descent dynamics (rescaled in time by $\gamma$) for the minimization problem (2.2). In the limit of $\gamma \to 0$, $v(t)$ approaches the eigenvector of the smallest eigenvalue instantly, and the GAD is reduced to the traditional eigenvector-following methodology. In this case, $v$ can be viewed as a function $v(x)$. For finite $\gamma$, (2.1) couples the dynamics of $x$ and $v$ simultaneously and still preserves the convergence to saddle points.

In contrast to the flow for $v$, the dynamics equation (2.1a) for the position $x$, however, is not in the form of the steepest descent dynamics of any scalar function. To see this, denote the GAD force as $F(x)$:

$$F_i(x) \triangleq f_i(x) - 2\sum_k v_i f_k(x) v_k,$$

where $f(x) \triangleq -\nabla V(x)$. It is easy to see that $\frac{\partial F_i}{\partial x_j} = -H_{ij} + 2v_i \sum_k H_{kj} v_k = -H + 2vv^{\mathsf{T}}H$, while its transpose $\frac{\partial F_j}{\partial x_i} = -H + 2Hvv^{\mathsf{T}}$. The necessary condition for the dynamics of $x$ being of a gradient type is that the Hessian $H$ commutes with the rank-1 matrix $vv^{\mathsf{T}}$, which generally does not hold since $v$ may not be the exact eigenvector of $H$ in the GAD. Even in the $\gamma \to 0$ limit where $v = v(x)$ is indeed the eigenvector of $H(x)$, the Jacobian matrix for $F_i(x) = f_i(x) - 2\sum_k v_i(x) f_k(x) v_k(x)$ is still not symmetric.

Therefore, the GAD is not as simple as a steepest descent flow, and no underlying energy function seems to exist to drive this dynamics. In the next section, we shall show that the GAD can be approximated by a steepest descent flow of a new objective function which is locally constructed. This is our iterative minimization formulation.

**3. The iterative minimization formulation.** In this section, we discuss how to define a new objective function to drive the system toward an index-1 saddle point

of the original energy function. The intuitive idea is to change the sign of the energy function $V$ along some direction, rather than reversing the direction of the force as in the GAD. The resulting Hessian then changes the sign of the smallest eigenvalue while keeping the other eigenvalues the same.

**3.1. The iterative scheme.** The framework we start with is the following iterative expression for $k = 0, 1, 2, \ldots$:

$$\begin{cases} v^{(k+1)} = \underset{\|u\|=1}{\operatorname{argmin}}\ u^{\mathsf{T}} H(x^{(k)}) u, & \text{(3.1a)} \\ x^{(k+1)} = \underset{y \in \mathcal{U}(x^{(k)})}{\operatorname{argmin}}\ \left( V(y) + W^{(k)}(y) \right), & \text{(3.1b)} \end{cases}$$

where $W^{(k)}$ is an unknown function to be determined. We need to construct $W^{(k)}$ such that $x^{(k)}$ converges to a saddle point of $V$. In (3.1b), $\mathcal{U}(x^{(k)})$ means a local neighborhood at $x^{(k)}$ (we refer the reader to Theorem 3.1(iii) next for a description of such a neighbourhood). We add this particular neighbor to highlight that the solution of the optimization (3.1b) we want is a local one. More specifically, (3.1b) has to be solved by using the special initial guess $y_{init} = x^{(k)}$.

The following two choices of the function $W^{(k)}$ serve our purpose:

$$W_1^{(k)}(y) = W_1(y; x^{(k)}, v^{(k+1)}), \qquad W_2^{(k)}(y) = W_2(y; x^{(k)}, v^{(k+1)}), \tag{3.2}$$

where, with the abuse of notation, we define

$$W_1(y; x, v) \triangleq -2V(y) + 2V\left(y - vv^{\mathsf{T}}(y - x)\right), \tag{3.3}$$

$$W_2(y; x, v) \triangleq -2V\left(x + vv^{\mathsf{T}}(y - x)\right), \tag{3.4}$$

which are two $\mathbb{R}^d \to \mathbb{R}$ functions parametrized by the position $x$ and the normalized direction $v$. Therefore, the new objective function $V + W$ depends on the current position $x$ and the direction $v$. In (3.2) for the choice of $W^{(k)}$, the direction $v^{(k+1)}$ is computed from the given $x^{(k)}$ by (3.1a). Therefore, (3.1) is actually an iterative scheme mapping $x^{(k)}$ to $x^{(k+1)}$ via a direction $v^{(k+1)}$.

Given a position $x$ and a direction $v$, we then have an affine hyperplane, denoted as $\mathcal{P}_{x,v}$, passing through the position $x$ with the normal $v$, i.e., $\mathcal{P}_{x,v} = \{y : (y-x)^{\mathsf{T}} v = 0\}$. Introduce the projection matrix $\Pi_v$ and $\Pi_v^{\perp} = \mathrm{I} - \Pi_v$, where

$$\Pi_v u = vv^{\mathsf{T}} u.$$

Then, the argument in the second term of $W_1$ is the point

$$y - vv^{\mathsf{T}}(y - x) = x + \Pi_v^{\perp}(y - x),$$

which is the projection of the point $y$ on the affine hyperplane $\mathcal{P}_{x,v}$. The position in $W_2$, $x + vv^{\mathsf{T}}(y - x) = x + \Pi_v(y - x)$, is the projection on the ray at $x$ with the direction $v$.

The intuition of the definitions of $W_1$ and $W_2$ is the following: If $y$ lies on the ray along $v$, then $W_2 = -2V(y)$. Consequently, the new energy function $V(y)+W_2(y; x, v)$ is to modify the potential $V(y)$ by reversing the sign of $V$ in the direction $v$. The choice of $V(y)+W_1(y; x, v)$, which is equal to $-V(y)+2V(x+\Pi^{\perp}(y-x))$, can be viewed as the reverse of the sign of $-V$ (instead of $V$) on the $(d-1)$-dimensional affine plane $\mathcal{P}_{x,v}$. Let us take a simple example of the quadratic function $V(y) = \frac{1}{2}\sum_{i=1}^{d} \mu_i y_i^2$,

where $\mu_1 < 0 < \mu_2 < \cdots < \mu_d$. The zero vector is the index-1 saddle point of $V$. $v = (1, 0, \ldots, 0)$ is the eigenvector corresponding to the smallest eigenvalue $\mu_1$. Then,

$$V(y) + W_1(y; x, v) = \mu_1 x_1^2 - \frac{1}{2}\mu_1 y_1^2 + \frac{1}{2}\sum_{i=2}^{d} \mu_i y_i^2,$$

$$V(y) + W_2(y; x, v) = -\sum_{i=2}^{d} \mu_i x_i^2 - \frac{1}{2}\mu_1 y_1^2 + \frac{1}{2}\sum_{i=2}^{d} \mu_i y_i^2.$$

The difference of these two functions of $y$ is just a constant $2V(x)$. Both are the convex quadratic functions of $y$, and they share the same Hessian $\mathrm{diag}\{-\mu_1, \mu_2, \ldots, \mu_d\}$ as well as the same minimizer 0, which is exactly the saddle point of $V$. So for any initial position $x^{(0)}$, the next iteration $x^{(1)}$ is the true solution. For the general function $V$, neither $V+W_1$ nor $V+W_2$ would be globally quadratic, and there are possibly multiple (local) minimizers. However, in the following, we show that it is always possible to define a meaningful local minimizer when the initial $x^{(0)}$ is sufficiently close to the saddle point.

**3.2. Convergence result.** We can formulate the saddle search problem as a fixed point problem in the iterative scheme (3.1) together with the defined $W_1$ and $W_2$ in (3.3) and (3.4). In fact, the function $W^{(k)}$ in the iterative scheme (3.1) can be some linear combination of $W_1$ and $W_2$ to achieve our purpose, too. In addition, the constant 2 showing in $W_1$ and $W_2$ can be relaxed. In the next theorem, we shall consider this general case to define the mapping from $x^{(k)}$ to $x^{(k+1)}$. Denote this mapping for the iteration as $\Phi(x)$. We shall show that the Jacobian matrix of $\Phi$ vanishes at the index-1 saddle point. This implies that the iterative scheme is of quadratic convergence.

THEOREM 3.1. *Assume that $V(x) \in \mathcal{C}^3(\mathbb{R}^d; \mathbb{R})$. For each $x$, let $v(x)$ be the normalized eigenvector corresponding to the smallest eigenvalue of the Hessian matrix $H(x) = \nabla^2 V(x)$, i.e.,*

$$v(x) = \operatorname*{argmin}_{u \in \mathbb{R}^d,\, \|u\|=1} u^T H(x) u.$$

*Given two real numbers $\alpha$ and $\beta$ satisfying $\alpha + \beta > 1$, we define the following function of the variable $y$:*

$$\begin{aligned} L(y; x, \alpha, \beta) = (1-\alpha)V(y) &+ \alpha V\Big(y - v(x)v(x)^T(y-x)\Big) \\ &- \beta V\Big(x + v(x)v(x)^T(y-x)\Big). \end{aligned} \tag{3.5}$$

*Suppose that $x^*$ is an index-1 saddle point of the function $V(x)$, i.e, $\nabla V(x^*)$ has only one negative eigenvalue $\lambda(x^*)$. Then the following statements are true.*

(i) *$x^*$ is a local minimizer of $L(y; x^*, \alpha, \beta)$.*

(ii) *There exists a neighborhood $\mathcal{U}$ of $x^*$ such that for any $x \in \mathcal{U}$, $L(y; x, \alpha, \beta)$ is strictly convex in $y \in \mathcal{U}$ and thus has a unique minimum in $\mathcal{U}$.*

(iii) *Define the mapping $\Phi : x \in \mathcal{U} \to \Phi(x) \in \mathcal{U}$, where $\Phi(x)$ is the unique local minimizer of $L$ in $\mathcal{U}$ for any $x \in \mathcal{U}$. Further assume that $\mathcal{U}$ contains no other stationary point of $V$ except $x^*$. Then the mapping $\Phi$ has only one fixed point $x^*$.*

(iv) $\Phi(x)$ *is differentiable in* $\mathcal{U}$. *The derivative of* $\Phi$ *vanishes at* $x^*$, *i.e., the Jacobian matrix*

$$\Phi_x(x^*) = 0. \tag{3.6}$$

*Proof. Part* (i): We calculate the first and second derivatives of $L(y; x, \alpha, \beta)$ with respect to $y$. The first order derivative is

$$\begin{aligned} \nabla_y L = (1-\alpha)\nabla V(y) + \alpha(I - vv^{\mathsf{T}})\nabla V\left(y - vv^{\mathsf{T}}(y-x)\right) \\ -\beta vv^{\mathsf{T}}\nabla V\left(x + vv^{\mathsf{T}}(y-x)\right). \end{aligned} \tag{3.7}$$

So, it is clear that $\nabla_y L(x^*; x^*, \alpha, \beta) = 0$ for any constants $\alpha$ and $\beta$ since $\nabla V(x^*) = 0$.

The Hessian matrix of $L$ is

$$\begin{aligned} \nabla_y^2 L(y; x, \alpha, \beta) =& (1-\alpha)\nabla^2 V(y) + \alpha(I - vv^{\mathsf{T}})\nabla^2 V(y - vv^{\mathsf{T}}(y-x))(I - vv^{\mathsf{T}}) \\ & - \beta vv^{\mathsf{T}}\nabla^2 V(x + vv^{\mathsf{T}}(y-x))vv^{\mathsf{T}}, \end{aligned} \tag{3.8}$$

which is simplified at $y = x^*$ and $x = x^*$ as follows:

$$\nabla_y^2 L(x^*; x^*, \alpha, \beta) = H(x^*) - (\alpha+\beta)\lambda(x^*)v(x^*)v(x^*)^{\mathsf{T}},$$

where the fact $H(x^*)v(x^*) = \lambda(x^*)v(x^*)$ is applied. Since $\lambda(x^*) < 0$, it follows that $\nabla_y^2 L(x^*; x^*, \alpha, \beta)$ is (strictly) positive definite if $\alpha + \beta > 1$.

*Part* (ii): The assumption that $V \in \mathcal{C}^3$ and $x^*$ is the index-1 saddle point of $V$ implies that the eigendirection $v(x)$ is continuously differentiable at $x^*$. Then (3.8), together with the continuity of $\nabla^2 V(x)$ and $v(x)$ at $x^*$, implies that the Hessian, $\nabla_y^2 L(y; x, v(x))$, which is now treated as a function of two variables $y$ and $x$, is continuously differentiable at $(y, x) = (x^*, x^*)$. In Part (i), we proved that at the point $(x^*, x^*)$, the Hessian is positive-definite as $\alpha + \beta > 1$. Thus, there exists a neighborhood of $(x^*, x^*)$, denoted as $\mathcal{N}$, such that the Hessian $\nabla_y^2 L(y; x)$ at each $(y, x) \in \mathcal{N}$ is still positive-definite. Select a neighborhood in the form of the product of two convex sets $\mathcal{U} \times \mathcal{U}$ inside the $2d$-dimensional set $\mathcal{N}$; then the set $\mathcal{U}$ is the desired one.

*Part* (iii): Suppose that there is a second fixed point $\hat{x} \in \mathcal{U}$ such that $\Phi(\hat{x}) = \hat{x}$. Then $\nabla_y L(\hat{x}; \hat{x}) = 0$. From (3.7),

$$\nabla_y L(\hat{x}; \hat{x}) = \nabla V(\hat{x}) - (\alpha+\beta)v(\hat{x})v(\hat{x})^{\mathsf{T}}\nabla V(\hat{x}) = 0.$$

Since $\alpha + \beta \neq 1$, then $\nabla V(\hat{x}) = 0$. But there is only one stationary point $x^*$ in $\mathcal{U}$, so $\hat{x}$ has to be the point $x^*$.

*Part* (iv): For each $x_0 \in \mathcal{U}$, $(\Phi(x_0), x_0)$ is the solution of the first order equation $\nabla_y L(\Phi(x_0); x_0) = 0$. It is clear that $\nabla_y L(y; x)$ is continuously differentiable at all $(y, x)$ in $\mathcal{U} \times \mathcal{U}$. In addition, $\nabla_y^2 L(y; x)$ is strictly positive-definite from Part (ii), thus nondegenerate in $\mathcal{U} \times \mathcal{U}$. Therefore, the implicit function theorem implies that $\Phi(x)$ is Lipschitz continuous and differentiable near $x_0$.

Next, we calculate the derivative of the mapping $\Phi$, denoted as $\Phi_x(x)$. For each $x \in \mathcal{U}$, $\Phi(x)$ is a solution of the first order equation (3.7). Thus the following equation holds:

$$(1-\alpha)\nabla V(\Phi(x)) + \alpha\left(I - v(x)v(x)^{\mathsf{T}}\right)\nabla V(\varphi_1(x)) - \beta\, v(x)v(x)^{\mathsf{T}}\nabla V(\varphi_2(x)) = 0, \tag{3.9}$$

where

$$\varphi_1(x) = \Phi(x) - v(x)v(x)^{\mathsf{T}}(\Phi(x) - x), \qquad \varphi_2(x) = x + v(x)v(x)^{\mathsf{T}}(\Phi(x) - x).$$

Now we take the derivative with respect to $x$ on both sides of (3.9); then

$$\begin{aligned}(1-\alpha)H(\Phi)\Phi_x + \alpha\left(I - vv^{\mathsf{T}}\right)H(\varphi_1)\varphi_{1,x} - \alpha v^{\mathsf{T}}\nabla V(\varphi_1)J - \alpha v\nabla V(\varphi_1)^{\mathsf{T}}J \\ = \beta vv^{\mathsf{T}}H(\varphi_2)\varphi_{2,x} + \beta v^{\mathsf{T}}\nabla V(\varphi_2)J + \beta v\nabla V(\varphi_2)^{\mathsf{T}}J,\end{aligned} \tag{3.10}$$

where the derivatives $J = \frac{\partial v(x)}{\partial x}$ and $\Phi_x = \frac{\partial \Phi}{\partial x}$ are the Jacobian matrix of $v(x)$ and $\Phi(x)$, respectively. The derivatives $\varphi_{1,x}$, $\varphi_{2,x}$ are defined similarly.

Note that $\Phi(x^*) = x^*$; thus $\varphi_1(x^*) = \varphi_2(x^*) = x^*$. Consequently $\nabla V(\varphi_1)$ and $\nabla V(\varphi_2)$ both vanish at $x^*$ since $\nabla V(x^*) = 0$. Meanwhile, since

$$\begin{aligned}\varphi_{1,x} &= \Phi_x - vv^{\mathsf{T}}(\Phi_x - I) - v^{\mathsf{T}}(\Phi - x)J - v(\Phi - x)^{\mathsf{T}}J, \\ \varphi_{2,x} &= I + vv^{\mathsf{T}}(\Phi_x - I) + v^{\mathsf{T}}(\Phi - x)J + v(\Phi - x)^{\mathsf{T}}J,\end{aligned}$$

then in particular at $x = x^*$, we have

$$\begin{aligned}\varphi_{1,x}(x^*) &= \left(I - v(x^*)v(x^*)^{\mathsf{T}}\right)\Phi_x(x^*) + v(x^*)v(x^*)^{\mathsf{T}}, \\ \varphi_{2,x}(x^*) &= I - v(x^*)v(x^*)^{\mathsf{T}} + v(x)v(x)^{\mathsf{T}}\Phi_x(x^*).\end{aligned}$$

Therefore, by noting $\nabla V(x^*) = 0$ again, (3.10) at $x = x^*$ becomes

$$\begin{aligned}&(1-\alpha)H(x^*)\Phi_x(x^*) \\ &+ \alpha\left(I - v(x^*)v(x^*)^{\mathsf{T}}\right)H(x^*)\left(I - v(x^*)v(x^*)^{\mathsf{T}}\right)\Phi_x(x^*) \\ &+ \alpha\left(I - v(x^*)v(x^*)^{\mathsf{T}}\right)\nabla^2 V(x^*)v(x^*)v(x^*)^{\mathsf{T}} \\ &- \beta v(x^*)v(x^*)^{\mathsf{T}}H(x^*)\left(I - v(x^*)v(x^*)^{\mathsf{T}}\right) \\ &- \beta v(x^*)v(x^*)^{\mathsf{T}}H(x^*)v(x^*)v(x^*)^{\mathsf{T}}\Phi_x(x^*) \\ &= 0.\end{aligned} \tag{3.11}$$

As $v(x)$ is the eigenvector of the Hessian $\nabla^2 V(x)$, i.e., $H(x)v(x) = \lambda(x)v(x)$, it is easy to verify that for each $x$, $(I - v(x)v(x)^{\mathsf{T}})\nabla^2 V(x)v(x)v(x)^{\mathsf{T}} = 0$ holds. Thus, any term in (3.11) without the Jacobian matrix $\Phi_x$ vanishes. So, (3.11) gives the following linear equation:

$$\left(H(x^*) - (\alpha+\beta)\lambda(x^*)v(x^*)v(x^*)^{\mathsf{T}}\right)\Phi_x(x^*) = 0, \tag{3.12}$$

which implies that $\Phi_x(x^*) = 0$ if and only if $\alpha + \beta \neq 1$. ∎

The above theorem implies the important property of the proposed iterative minimizing formulation in section 3.1 if the energy function $V$ has a higher regularity $\mathcal{C}^4$.

COROLLARY 3.2. *Under the assumptions of Theorem* 3.1*, assume that* $V(x) \in \mathcal{C}^4(\mathbb{R}^d;\mathbb{R})$*; then the iterative scheme* $x^{(k+1)} = \Phi(x^{(k)})$ *has exactly the second order (local) convergence rate.*

*Proof.* Since $V$ is $\mathcal{C}^4$, it follows that $\nabla_y L$ has the regularity $\mathcal{C}^2$ and is in the neighborhood $\mathcal{U}$. It follows that $\Phi(x)$ is continuously differentiable at $x^*$ based on the second order pseudoexpansion [2] for the first order equation $\nabla_y L = 0$.

Since $\Phi_x(x^*) = 0$, there is a neighborhood of $x^*$ such that $\|\Phi_x(x)\|$ is strictly less than 1 in this neighborhood. Thus, the local convergence comes from the contraction mapping principle.

The second order convergence rate is due to the fact that the Jacobian matrix $\Phi_x(x^*)$ vanishes. One can carry out further calculations and observe that the second

derivative of $\Phi(x)$ at $x^*$ does not trivially vanish. So, the iteration $x \to \Phi(x)$ locally converges to $x^*$ exactly at the quadratic rate. ∎

For the quadratic example in section 3.1, we have the following trivial result.

COROLLARY 3.3. *If $V(x) = \frac{1}{2}x^T Hx$, where $H$ is a constant symmetric matrix and has only one negative eigenvalue, then $\Phi(x) = 0$ for all $x$ when $\alpha, \beta$ in Theorem* 3.1 *satisfy $\alpha + \beta > 1$.*

We remark that the $\Phi(x)$ is well defined in a local neighborhood of $x^*$. In implementations, the local solution of the new objective function $L(y; x^{(k)})$ in (3.5) is pursued with the initial guess $y_0 = x^{(k)}$. This choice of the initial guess is not only very simple to pick up but also excludes the possibilities of overshooting to other local solutions which are not relevant to the saddle point of interest.

**3.3. Solve subproblem of minimization.** The iterative minimization formulation (3.1) consists of solving a subproblem of minimization at each iteration. Corollary 3.2 suggests that the quadratic convergence rate is achieved when the subproblem is solved exactly and the correct local minimizer is found. In practice, one may not need to solve the subproblem of the minimization exactly or with high accuracy, and the superlinear convergence might be achieved in certain circumstances. Many existing eigenvector-following methods like the dimer method could be viewed as some special discretization for the subproblem. We just present a result about the connection of the GAD and the iterative minimization formulation.

THEOREM 3.4. *Assume $x^{(k)}$ is near the index-*1 *saddle point $x^*$. Suppose that one solves the subproblem* $\min_y L(y; x^{(k)}, \alpha, \beta)$ *in Theorem* 3.1 *by only one single steep descent method with the step size $\delta t$,*

$$x^{(k+1)} = x^{(k)} - \delta t \, \nabla_y L(x^{(k)}; x^{(k)}, \alpha, \beta). \tag{3.13}$$

*Then the sequence $\{x^{(k)}\}$ is the discrete solution of the Euler method with the time step $\delta t$ for the following version of the GAD:*

$$\dot{x} = -\nabla V(x) + (\alpha + \beta)\Pi_{v(x)} \nabla V(x). \tag{3.14}$$

*Proof.* The conclusion is obvious by noting the following and the fact in (3.7):

$$\begin{aligned} x_{k+1} &= x_k - \delta t \, \nabla_y L(x_k; x_k, \alpha, \beta) \\ &= x_k - \delta t \big( (1-\alpha)\nabla V(x_k) + \alpha (I - v(x_k)v(x_k)^T \nabla V(x_k)) \\ &\qquad - \beta v(x_k) v(x_k)^T \nabla V(x_k) \big). \end{aligned}$$ ∎

*Remark* 1. If the subproblem for the direction $v$ is also solved by the steepest descent method, then $(x^{(k)}, v^{(k)})$ corresponds to the original version of the GAD (2.1).

The subproblem at each iteration consists of the minimization for the position and the direction. Some fast algorithms have been developed for solving the min-mode direction, in particular, where the Hessian is not explicitly available and the force is calculated from the first principle [17]. The new numerical challenge in implementing the iterative minimization formulation efficiently is the minimization of $L$ for the position to get new $x^{(k+1)}$. Of course, one is not limited to using the steepest descent method to solve this subproblem as in Theorem 3.4. For example, the conjugate gradient (CG) method can be applied with a certain level of tolerance. Details about these accelerating techniques will be postponed to a separate paper.

We now discuss the choice of two parameters $\alpha$ and $\beta$ in our formulation. Theoretically by Theorem 3.1, the condition that $\alpha + \beta > 1$ is sufficient for the algorithm

to achieve the local quadratic convergence. In practice, a better choice of $\alpha$ and $\beta$ may help reduce the condition number of the subproblem, which is the ratio of the maximum eigenvalue and the smallest eigenvalue of the Hessian $\nabla_y^2 L$. The calculation in the proof of Theorem 3.1 has shown that at the saddle point $x^*$, the eigenvalues of $\nabla_y^2 L$ are $(1-\alpha-\beta)\lambda_1, \lambda_2, \lambda_3, \ldots, \lambda_d$ where $\lambda_1 < 0 < \lambda_2 < \cdots < \lambda_d$ are eigenvalues of $\nabla^2 V(x^*)$. Hence, to minimize the condition number of $\nabla_y^2 L$, the optimal choice of $\alpha$ and $\beta$ needs to satisfy

$$1 + \frac{\lambda_2}{|\lambda_1|} \leq \alpha + \beta \leq 1 + \frac{\lambda_d}{|\lambda_1|},$$

and the resulting optimal condition number is $\lambda_d/\lambda_2$. In practice, a rough estimate of $\lambda_2$ may be used to select the parameter $\alpha + \beta$ at each iteration.

When the initial guess of the iterative method is in the convex region of the original energy function, for example, a local minimum, the function $L$ will have no lower bound locally and the minimization subproblem does not have a solution. One can handle this situation using the traditional techniques implemented in many eigenvector-following–type methods. One simple remedy is to seek the solution only within a ball or hypercube of a proper size near the current $x^{(k)}$. Such remedies are not necessarily needed when $\lambda_1$ is negative.

**4. Saddle points on manifold.** In some applications, the configuration of the system may be subject to one or more constraints, for instance, the conservation laws of some physical quantities. Suppose that these constraints realize a Riemann manifold $\mathcal{M}$ embedded in $\mathbb{R}^d$. The index-1 saddle point of the energy function restricted on $\mathcal{M}$ is still the transition state of interest. The calculation of the saddle point on the manifold calls for the attention to the constraints associated with the manifold. In this section, we want to extend the GAD and iterative minimization formulation onto the manifold $\mathcal{M}$. This goal can be easily achieved for the GAD by a simple projection procedure, as in [8, 26], but it requires extra work for the iterative minimization formulation.

We assume that the manifold $\mathcal{M}$ is characterized by $p$ (independent) equality constraints: $c_i(x) = 0$ for $i = 1, 2, \ldots, p$, where $c_i$ are $\mathbb{R}^d \to \mathbb{R}$ smooth functions. To maintain the right mix of abstraction and concreteness, we use the extrinsic variables $x$ in $\mathbb{R}^d$ for $\mathcal{M}$. The tangent space $\mathcal{T}_x$ at each point $x$ of the manifold $\mathcal{M}$ is thus the orthogonal complement in $\mathbb{R}^d$ to the normal space spanned by the gradients of the $p$ constraints, $span\{\nabla c_i(x), i = 1, 2, \ldots, p\}$. The concepts of the local minimum and the index-1 saddle point of the smooth energy function $V(x)$ can be extended to the manifold case without any difficulty [1]. We skip the rigorous math definitions since these concepts are quite intuitive.

We start with the calculation of the eigenvector $v$ in the tangent space $\mathcal{T}_x$ corresponding to the smallest eigenvalue of the (projected) Hessian matrix of the energy function $V$. This direction $v$ minimizes the Rayleigh quotient among all possible vectors in $\mathcal{T}_x$:

$$v = \operatorname*{argmin}_{\|u\|=1, u \in \mathcal{T}_x} u^{\mathsf{T}} \nabla^2 V(x) u,$$

or, equivalently,

$$v = \operatorname*{argmin}_{\|u\|=1, u \in \mathbb{R}^d} \Big\{ u^{\mathsf{T}} \nabla^2 V(x) u \Big| \langle \nabla c_i(x), u \rangle = 0 \ \forall i = 1, 2, \ldots, p \Big\}. \tag{4.1}$$

The steepest descent flow of this constrained minimization problem is

$$\gamma\dot{v} = -\Pi_{\mathcal{T}_x}\left[\nabla^2 V(x)v\right] + \eta v,$$

where $\Pi_{\mathcal{T}_x}$ is the orthogonal projection of $\mathbb{R}^d$ to the vector space $\mathcal{T}_x$ and the scalar $\eta = \left\langle \Pi_{\mathcal{T}_x}\left[\nabla^2 V(x)v\right], v\right\rangle$ is to enforce the unit length of $v$. Many existing fast algorithms for the original rotation step to solve $v$ in $\mathbb{R}^d$ can be readily modified for the constrained problem (4.1).

Next, we discuss the dynamics or the iterations for the position variable $x$. For the dynamics of $x$ in GAD, we can simply project the GAD force $(-\mathrm{I} + 2vv^{\mathsf{T}})\nabla V(x)$ onto the tangent space $\mathcal{T}_x$, i.e.,

$$\dot{x} = \Pi_{\mathcal{T}_x}\left[(-\mathrm{I} + 2vv^{\mathsf{T}})\nabla V(x)\right];$$

then the trajectory of the GAD stays on the manifold $\mathcal{M}$ all the time. However, for the iterative minimization formulation (3.1), the need of projection on $\mathcal{M}$ complicates our discussion. Specifically, for a given $x \in \mathcal{M}$ and $v \in \mathcal{T}_x$, one must find a geodesic curve on $\mathcal{M}$ by following the geodesic flow which can be described in terms of these constraints functions $c_i(x)$ [3]. Let $\xi(s)$ ($s \in \mathbb{R}$) be the geodesic curve satisfying $\xi(0) = x$ and $\xi'(0) = v$. For each point $y \in \mathcal{M}$ near $x$, under some mild condition, we can define the projection of $y$ onto the geodesic $\xi$ as $\xi(s_y)$, where $s_y \triangleq \operatorname{argmin}_s \operatorname{dist}(\xi(s), y)$. Here "dist" is the distance between two points of the manifold $\mathcal{M}$: the infimum of the lengths of all continuously differentiable curves on $\mathcal{M}$ joining these two points. The argument of the $W_2$ function is then the point which has minimal distance on $\mathcal{M}$ to the curve $\xi(s)$, i.e., the "projection" of $y$ to $\xi$. Therefore, the formula of $W_2$ on $\mathcal{M}$ is

$$W_2(y) = -2V(\xi(s_y)). \tag{4.2}$$

In principle, the same strategy can be applied for the $W_1$ function where one should use the minimal distance to the set of geodesic curves whose tangents are in $\mathcal{T}_x$ but orthogonal to $v$.

In a nutshell, the iterative minimization scheme on the manifold $\mathcal{M}$ specified by the $p$ constraints $c_i(x) = 0$ can be written as follows:

$$\begin{cases} v^{(k+1)} = \displaystyle\operatorname*{argmin}_{\|u\|=1, u\in\mathbb{R}^d} \left\{ u^{\mathsf{T}}\nabla^2 V(x^{(k+1)})u \;\middle|\; \left\langle \nabla c_i(x^{(k)}), u \right\rangle = 0 \ \forall i \right\}, & \text{(4.3a)} \\ x^{(k+1)} = \displaystyle\operatorname*{argmin}_{y\in\mathcal{U}(x^{(k)})} \left\{ V(y) + W^{(k)}(y) \;\middle|\; c_i(y) = 0 \ \forall i = 1, 2, \ldots, p \right\}, & \text{(4.3b)} \end{cases}$$

where $W^{(k)}$ (which depends on $x^{(k)}$ and $v^{(k+1)}$) is defined through the above-mentioned $W_1$, $W_2$, or their linear combination in the same way as in Theorem 3.1. To illustrate the above idea, the example of the sphere $S^2$ in $\mathbb{R}^3$ is presented in section 6.3. The numerical result for a quadratic energy function on this manifold shows that the iterative scheme (4.3) also has the quadratic convergence rate.

**5. Saddle with higher index.** Reference [11] about the GAD has extended from the index-1 saddle point to saddle points of index $m$ for the $m > 1$ case with the help of a dilation technique. Our new iterative minimization formulation proposed above for the index-1 saddle can also be extended to the case of saddles with index greater than 1. Suppose that one has found $m$ eigenvectors of the Hessian matrix $\nabla^2 V(x)$, $v_1, v_2, \ldots, v_m$, corresponding to the $m$ smallest eigenvalues, respectively. We denote $S$ as the set of all subsets of $\{1, \ldots, m\}$ except the empty set. For every $s \in S$,

we have $s = \{i_1, \dots, i_k\} \subset \{1, \dots, m\}$ with $k \le m$ and $1 \le i_1 < \cdots < i_k \le m$ . The projection onto the plane spanned by $k$ column vectors $\{v_{i_1}, v_{i_2}, \dots, v_{i_k}\}$ is associated with the following matrix:

$$\Pi_s = V_s V_s^{\mathsf{T}} = \begin{bmatrix} v_{i_1} & v_{i_2} & \cdots & v_{i_k} \end{bmatrix} \begin{bmatrix} v_{i_1}^{\mathsf{T}} \\ v_{i_2}^{\mathsf{T}} \\ \vdots \\ v_{i_k}^{\mathsf{T}} \end{bmatrix}.$$

Let $\Pi_s^{\perp} = \mathrm{I} - \Pi_s$. The objective function for the subproblem, which is a generalization to (3.5), is now given by

$$\begin{aligned} L(y; x, \alpha, \beta) = \left(1 - \sum_{s \in S} \alpha_s\right) V(y) + \sum_{s \in S} \alpha_s V(x + \Pi_s^{\perp}(y - x)) \\ - \sum_{s \in S} \beta_s V(x + \Pi_s(y - x)), \end{aligned} \tag{5.1}$$

where $\alpha = (\alpha_s)_{s \in S}$ and $\beta = (\beta_s)_{s \in S}$. For example, the function (5.1) in the index-2 case is

$$\begin{aligned} L(y; x, \alpha, \beta) &= (1 - \alpha_1 - \alpha_2 - \alpha_{1,2}) V(y) \\ &\quad + \alpha_1 V\left(x + \Pi_1^{\perp}(y - x)\right) + \alpha_2 V\left(x + \Pi_2^{\perp}(y - x)\right) + \alpha_{1,2} V\left(x + \Pi_{1,2}^{\perp}(y - x)\right) \\ &\quad - \beta_1 V\left(x + \Pi_1(y - x)\right) - \beta_2 V\left(x + \Pi_2(y - x)\right) - \beta_{1,2} V\left(x + \Pi_{1,2}(y - x)\right). \end{aligned}$$

In parallel to Theorem 3.1, we have the following theorem for the index-$m$ case. Its proof is similar to the proof of Theorem 3.1 but technically lengthier and thus is skipped.

THEOREM 5.1. *Assume that $V(x) \in \mathcal{C}^4(\mathbb{R}^d; \mathbb{R})$. For each $x$, let $v_1(x), \dots, v_m(x)$ be $m$ normalized eigenvectors corresponding to the smallest eigenvalues of the Hessian matrix $H(x) = \nabla^2 V(x)$, i.e.,*

$$[v_1(x), \dots, v_m(x)] = \operatorname*{argmin}_{U = [u_1, \dots, u_m],\, U^T U = I} \operatorname{trace} U^T \nabla^2 V(x) U.$$

*The function $L(y; x, \alpha, \beta)$ of the variable $y$ is defined as in (5.1), and it is assumed that*

$$\sum_{s \in S} (\alpha_s + \beta_s) > 1.$$

*Suppose that $x^*$ is an index-$m$ saddle point of the function $V(x)$. Then the following statements are true.*

(i) *$x^*$ is a local minimizer of $L(y; x^*, \alpha, \beta)$.*

(ii) *There exists a neighborhood $\mathcal{U}$ of $x^*$ such that for any $x \in \mathcal{U}$, $L(y; x, \alpha, \beta)$ is strictly convex in $y \in \mathcal{U}$ and thus has a unique minimum in $\mathcal{U}$. We define $\Phi(x)$ to be this minimizer for the given $x$.*

(iii) *The mapping $\Phi$ has only one unique fixed point $x^*$ in $\mathcal{U}$.*

(iv) *The mapping $\Phi$ is differential in $\mathcal{U}$. The Jacobian matrix of $\Phi$ vanishes at $x^*$; i.e.,*

$$\Phi_x(x^*) = 0.$$

As a consequence of the above theorem, the iterative scheme

$$\begin{cases} x^{(k+1)} = \operatorname{argmin}_y L(y; x^{(k)}, \alpha, \beta), \\ [v_1^{(k+1)}, \ldots, v_m^{(k+1)}] = \operatorname{argmin}_{U=[u_1,\ldots,u_m],\, U^\top U = I} \operatorname{trace} U^\top \nabla^2 V(x^{(k)}) U \end{cases}$$

converges to the index-$m$ saddle point $x^*$ quadratically if the starting point $x^{(0)}$ is close enough to $x^*$.

## 6. Examples.

**6.1. A simple two-dimensional example.** First, we review a two-dimensional example from [11]:

$$V(x,y) = \frac{1}{4}(x^2-1)^2 + \frac{1}{2}\mu y^2,$$

where $\mu$ is a positive parameter. For this system, $x_\pm = (\pm 1, 0)$ are two local minima, and $(0,0)$ is the index-1 saddle point. The eigenvalues and eigenvectors of the Hessian matrix at a point $(x,y)$ are

$$\begin{aligned} &\lambda_1 = 3x^2 - 1 \quad \text{and} \quad v_1 = (1,0), \\ &\lambda_2 = \mu \quad \text{and} \quad v_2 = (0,1). \end{aligned}$$

Note that when $|x| \le \sqrt{\frac{1+\mu}{3}}$, $\lambda_1 \le 0 < \lambda_2$. The min-mode is $v_1$ if $|x| < \sqrt{\frac{1+\mu}{3}}$ and is $v_2$ if $|x| > \sqrt{\frac{1+\mu}{3}}$.

Suppose that at iteration $k$, the position is $(x_k, y_k)$. Then, the modified objective functions $V + W_1$ and $V + W_2$ in the iterative minimization formulation are defined as follows:

$$\begin{cases} V + W_1^{(k)} = -\frac{1}{4}(x^2-1)^2 + \frac{1}{2}\mu y^2 + \frac{1}{2}(x_k^2-1)^2 & \text{if } |x| < \sqrt{\frac{1+\mu}{3}}, \\ V + W_1^{(k)} = \frac{1}{4}(x^2-1)^2 - \frac{1}{2}\mu y^2 + \mu y_k^2 & \text{if } |x| > \sqrt{\frac{1+\mu}{3}}, \end{cases}$$

$$\begin{cases} V + W_2^{(k)} = -\frac{1}{4}(x^2-1)^2 + \frac{1}{2}\mu y^2 - \mu y_k^2 & \text{if } |x| \le \sqrt{\frac{1+\mu}{3}}, \\ V + W_2^{(k)} = \frac{1}{4}(x^2-1)^2 - \frac{1}{2}\mu y^2 - \frac{1}{2}(x_k^2-1)^2 & \text{if } |x| > \sqrt{\frac{1+\mu}{3}}. \end{cases}$$

These are piecewise continuous functions, and the difference between $W_1$ and $W_2$ is only a constant. In the domain where $|x| < \min(1, \sqrt{\frac{1+\mu}{3}})$, the original saddle $(0,0)$ is the unique interior minimal point. Outside of this domain, the modified function $V + W_1$ or $V + W_2$ has no lower bound. So, the iterative minimizing formulation works only when the initial guess satisfies $|x| < \sqrt{\frac{1+\mu}{3}}$.

**6.2. The three-hole example.** In this example, we study a two-dimensional energy function from [22, 19], where there are three local minima. The formula of this energy function is

$$\begin{aligned} V(x,y) = {} & 3e^{-x^2-(y-\frac{1}{3})^2} - 3e^{-x^2-(y-\frac{5}{3})^2} - 5e^{(x-1)^2-y^2} - 5e^{(x+1)^2-y^2} \\ & + 0.2x^4 + 0.2\left(y - \frac{1}{3}\right)^4. \end{aligned}$$

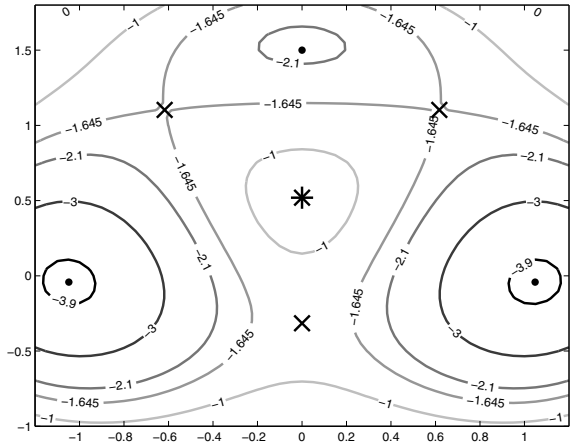


FIG. 1. *The three-hole potential: Three minima (black dots) approximately at* $(\pm 1, 0)$ *and* $(0, 1.5)$*, a maximum* $(*)$ *at* $(0, 0.5)$*, and three saddle points* $(\times)$ *at* $(0, -0.31582)$ *and* $(\pm 0.61727, 1.10273)$.

We refer the reader to Figure 1 for its contour plot.

SP1 $= (0, -0.31582)$ and SP2 $= (-0.61727, 1.10273)$ are the two saddles of interest. We first demonstrate the quadratic convergence when the initial guess is near the saddle points. Table 1 shows the errors at each iteration, from which it is observed that the iterative scheme has the quadratic convergence rate.

TABLE 1

*Errors of six runs with random initial guesses on the circle of radius* 0.2 *with the center at the target saddle point. Different values of* $(\alpha, \beta)$ *for the modified objective functions in the subproblem are shown in parentheses in the first row of the table. The three runs on the left converge to SP*1*, and the three runs on the right converge to SP*2*.*

| Iter | $(2, 0)$ | $(0, 2)$ | $(1, 1)$ | $(2, 0)$ | $(0, 2)$ | $(1, 1)$ |
|---|---|---|---|---|---|---|
| 1 | 5.042e-002 | 2.979e-002 | 2.924e-002 | 1.672e-002 | 3.024e-002 | 4.342e-002 |
| 2 | 1.376e-005 | 5.470e-004 | 1.671e-004 | 9.327e-006 | 3.445e-004 | 3.194e-004 |
| 3 | 7.245e-011 | 2.573e-008 | 2.434e-008 | 2.527e-011 | 1.116e-008 | 1.233e-008 |
| 4 | 5.023e-016 | 5.551e-016 | 3.951e-016 | 2.482e-016 | 3.886e-016 | 4.965e-016 |

If the initial guess is close to the local minima of $V$, then the Hessian of $V$ at the initial point is positive-definite, while the modified objective function $L$ has one negative eigenvalue and has no lower bound. As discussed in section 3.3, we set a maximum step size 0.25 in both the $x$- and the $y$-direction at each iteration to maintain the stability at this initial stage. Table 2 shows the result for initial points which are 0.1 away from $(-1, 0)$, one of the two deep minima. Some runs converge to SP2, and others converge to SP1. It is observed that at the first few steps, the decreasing of the errors is slow, but when it approaches the saddle, the smallest eigenvalue of the Hessian becomes negative, and it follows that the iterative minimization method starts to show quadratic convergence.

We also test the effects of the inexact solution of the subproblem. In solving the subproblem by the CG method, we perform only three steps of the CG iteration. The initial guesses are chosen on the circle centered at the saddle points with 0.2 radius. The results are shown in Table 3 with two different parameter sets ($\alpha = 2, \beta = 0$ and

TABLE 2

*Errors of six random runs with initial guesses on the circle of radius* 0.1 *with the center at the local minimum* $(-1, 0)$. *The three runs on the left converge to SP*1*, and the three runs on the right converge to SP*2.

| Iter | $(2,0)$ | $(0,2)$ | $(1,1)$ | $(2,0)$ | $(0,2)$ | $(1,1)$ |
|---|---|---|---|---|---|---|
| 1 | 8.434e-001 | 8.568e-001 | 8.496e-001 | 1.160e+000 | 1.170e+000 | 1.262e+000 |
| 2 | 6.891e-001 | 7.026e-001 | 6.954e-001 | 9.922e-001 | 9.970e-001 | 1.077e+000 |
| 3 | 5.731e-001 | 5.858e-001 | 5.790e-001 | 8.512e-001 | 8.537e-001 | 9.464e-001 |
| 4 | 5.216e-001 | 5.317e-001 | 5.263e-001 | 6.983e-001 | 6.939e-001 | 8.020e-001 |
| 5 | 3.862e-001 | 3.848e-001 | 3.841e-001 | 5.273e-001 | 5.028e-001 | 6.397e-001 |
| 6 | 1.776e-001 | 1.740e-001 | 1.742e-001 | 3.391e-001 | 3.030e-001 | 4.542e-001 |
| 7 | 1.983e-002 | 4.266e-002 | 1.987e-002 | 1.511e-001 | 1.291e-001 | 2.569e-001 |
| 8 | 7.314e-007 | 3.343e-004 | 2.834e-004 | 2.975e-002 | 1.207e-002 | 8.031e-002 |
| 9 | 3.756e-012 | 9.654e-009 | 4.088e-008 | 2.093e-006 | 2.879e-005 | 3.919e-003 |
| 10 | | 6.810e-016 | 7.773e-015 | 1.734e-015 | 1.912e-011 | 7.345e-006 |
| 11 | | | | | 3.140e-016 | 2.745e-011 |

TABLE 3

*Errors of four runs with random initial guesses on the circle of radius* 0.2 *with the center at the target saddle points. Use a three-step nonlinear CG method to solve the subproblem of minimization inexactly. The two runs on the left converge to SP*1*, and the two runs on the right converge to SP*2.

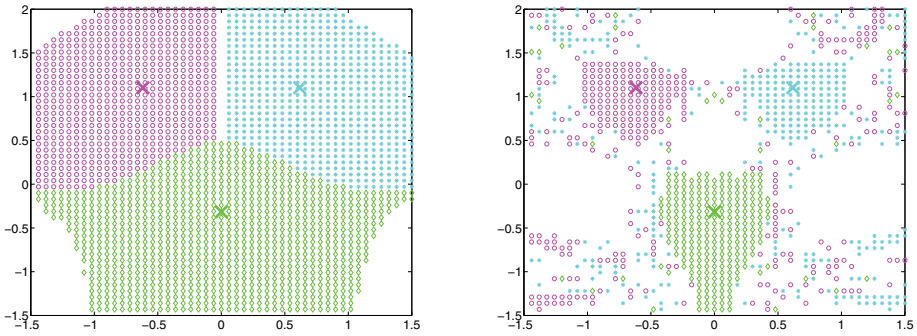
| Iter | $(2,0)$ | $(0,2)$ | $(2,0)$ | $(0,2)$ |
|---|---|---|---|---|
| 1 | 4.476e-02 | 2.723e-02 | 5.096e-02 | 1.877e-02 |
| 2 | 1.262e-04 | 3.256e-04 | 7.756e-05 | 8.386e-04 |
| 3 | 2.863e-08 | 6.047e-06 | 1.270e-09 | 4.176e-05 |
| 4 | 5.317e-13 | 3.316e-10 | 7.830e-13 | 1.8827e-09 |
| 5 | | 7.325e-13 | | 4.3853e-11 |

$\alpha = 0, \beta = 2$). In comparison to the case of an exact solution for the subproblem in Table 1, the efficiency of the algorithm is not affected much, and the local convergence rate is still quite close to the second order.

In the end, for this two-dimensional example, we plot the domain of attraction for our algorithm to compare with the performance of the Newton method. Note that our purpose here is to look for index-1 saddle points. We choose initial guesses from $50 \times 50$ grid points uniformly in the rectangular region $[-1.5, 1.5] \times [-1.5, 2.0]$. These points are labelled in Figure 2 by three different marks in three colors, according to which saddle point (shown by a cross symbol in the same color as that of its initial guesses) they converge to, respectively. The grid point is left blank in the case of no convergence. The figure demonstrates that our iterative minimization formulation (IMF) scheme has a larger (and continuous) domain of attraction for each saddle point than the Newton method.

**6.3. A quadratic function on the sphere $S^2$.** We illustrate the proposal in section 4 for the constrained problem by considering a simple example of $\mathcal{M} = S^2$ embedded in $\mathbb{R}^3$ on which a quadratic energy function $V(x_1, x_2, x_3) = x_1^2 + 2x_2^2 + 3x_3^2$ is defined. The constraint is that $x_1^2 + x_2^2 + x_3^2 = 1$. It is easy to verify that saddle points $(0, \pm 1, 0)$, minimizers $(\pm 1, 0, 0)$, and maximizers $(0, 0, \pm 1)$ of $V$ are generated due to this constraint.

As mentioned in section 4, for a given $x \in S^2$ and $v \in \mathcal{T}_x(S^2)$, the projection of a point $y$ on $S^2$ is associated with a geodesic curve $\xi(s)$. Here the geodesic $\xi$ is simply the great circle passing the point $x$ along the direction $v$. Thus $\xi$ can be written in the parametrized form $\xi(\theta) = x \cos\theta + v \sin\theta$. It follows that the geodesic distance $\mathrm{dist}(\xi(\theta), y) = \arccos\langle \xi(\theta), y \rangle$ achieves the minimum at $\theta = \theta_y$,



(a) IMF (b) the Newton method

FIG. 2. *Comparison of domains of attractions of saddle points for our IMF and the Newton method.*

where $\theta_y$ is equal to $\arctan \frac{\langle v,y\rangle}{\langle x,y\rangle}$ or $\arctan \frac{\langle v,y\rangle}{\langle x,y\rangle} + \pi$, depending on which value gives smaller distance. Then the projection point of $y$ is $x\cos\theta_y + v\sin\theta_y$. So we have $W_2(y) = -2V(x\cos\theta_y + v\sin\theta_y)$ for this $S^2$ example.

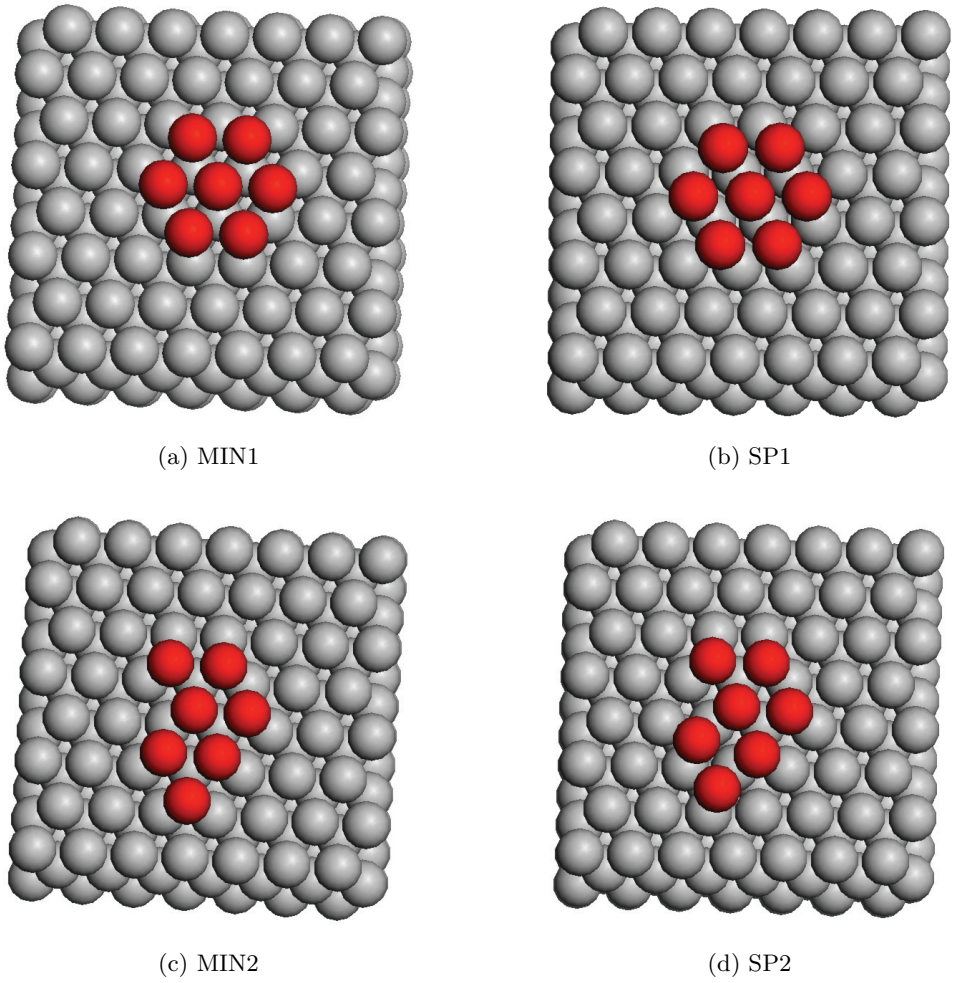
Next we also derive the $W_1$ expression for this $S^2$ case. Since the tangent space $\mathcal{T}_x(S^2)$ is two-dimensional, the orthogonal complement of $v$ in this space is spanned by just a single vector, denoted as $\tilde{v}$. It follows then that $W_1(y) = -2V(y)+2V(x\cos\tilde{\theta}_y + v\sin\tilde{\theta}_y)$, where $\tilde{\theta}_y$ is defined similarly to $\theta_y$ by replacing $v$ by $\tilde{v}$.

The numerical results based on the construction of the above $W_1$ and $W_2$ are presented in Table 4. The initial guess is 0.1 distance to the minimum point $(1, 0, 0)$. The numerical data of the errors between the solution and the true saddle point in this table again confirm the quadratic convergence rate. We remark that it is important to use the projection associated with the geodesic curve in the above construction of $W_1$ and $W_2$. One alternative idea might be to use the projection in the Euclidean space $\mathbb{R}^3$ as if there were no constraints, and then pull back to $S^2$. For instance, one may use the following: $W_2(y) = -2V(R_x(v^{\mathsf{T}}(y-x)v))$, where $R_x(u) = \frac{x+u}{\|x+u\|}$ is a retraction mapping the tangent space $\mathcal{T}_x$ to the sphere $S^2$. However, our numerical result for the same example here shows that this choice gives only a linear convergence rate. The missing curvature information of the manifold in this naive orthogonal projection approach seems to be the reason for lowering the convergence order.

TABLE 4
*Errors of $S^2$ example.*

| Iter | 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|---|
| $V + W_1$ | 1.3900e+00 | 2.3217e-01 | 1.4234e-03 | 2.0994e-08 | 1.7684e-15 |
| $V + W_2$ | 1.2902e+00 | 5.6506e-01 | 6.5923e-02 | 8.7484e-06 | 1.2433e-16 |

**6.4. An atomic model system.** This is an application of our method to the celebrated test problem of a 7-atom island on the (111) surface of an FCC metal [12]. In this example the structure has a 6-layer slab, each layer of which contains 56 atoms, and 7 atoms at the top of the slabs. The bottom three layers in the slab are frozen. There are in total $56 \times 3 + 7 = 175$ atoms that are free to move. All of the atoms in



(a) MIN1 (b) SP1

(c) MIN2 (d) SP2

FIG. 3. *The 7-atom island model. Two local minima and two saddle points are shown, denoted as MIN1, SP1, MIN2, SP2, respectively.*

this simulation are identical.

The interaction between the atoms is the simple pairwise additive Morse potential

$$V(R) = A[e^{-2a(R-R_0)} - 2e^{-a(R-R_0)}]$$

with parameters chosen to reproduce diffusion barriers on platinum surfaces ($A = 0.7102$eV, $a = 1.6047\mathring{A}^{-1}$, $R_0 = 2.8970\mathring{A}$). This potential is cut and shifted by V($R_C$), where $R_C = 9.5\mathring{A}$ is the cut-off distance. The minimum energy lattice constant $2.74412\mathring{A}$ is used.

In Figure 3 we show two local minima as well as two saddle points. All saddle points lead from the close packed heptamer (shown in red) to some adjacent state. We applied our iterative minimization method to this large-scale system. The initial guess of position is chosen near the minima, and the initial direction is randomly selected. The eigenvector corresponding to the minimum eigenvalue is solved by an

efficient method proposed in [17]. The maximum step size in the subproblem for the position is set as 0.2. In this implementation, we used the nonlinear CG method with the tolerance set to $10^{-16}$ so that the subproblem is solved accurately enough. The accuracy of each entry in the force at the numerical saddle points we found is between $10^{-10}$ and $10^{-11}$. The error is then defined as the Euclidean distance from the current position to the saddle points.

The numerical results are presented in Table 5. Since our initial guess is very close to the local minimum, it is not surprising that the first several iteration steps have a slow decay of the errors since the effect of following the eigenvector of the smallest eigenvalue has not kicked in. The fast convergence rate is observed as expected at the subsequent stage. In this example, the exact solver for each subproblem was used in Table 5; thus the computational overhead is large, compared with other algorithms requiring no subproblems to solve. We also introduced a simple inexact solver by limiting only two iterations of CG for the subproblem. The resulting convergence rate deteriorates due to inexact solver and linear convergence is observed. The right balance between the fast convergence rate and the large computational overhead requires a careful design of the tolerance in the inexact solver.

TABLE 5

*Errors of six runs with random initial guesses near the local minima as well as with different auxiliary potentials. The three runs on the left start from the initial guesses near MIN*1 *and converge to SP*1*. The three runs on the right start from the initial guesses near MIN*2 *and converge to SP*2*.*

| Iter | $V+W_1$ | $V+W_2$ | $\frac{2V+W_1+W_2}{2}$ | $V+W_1$ | $V+W_2$ | $\frac{2V+W_1+W_2}{2}$ |
|---|---|---|---|---|---|---|
| 1 | 2.014e+000 | 1.832e+000 | 1.803e+000 | 1.633e+000 | 1.695e+000 | 1.521e+000 |
| 2 | 1.837e+000 | 1.695e+000 | 1.760e+000 | 1.599e+000 | 1.575e+000 | 1.488e+000 |
| 3 | 1.729e+000 | 1.575e+000 | 1.693e+000 | 1.535e+000 | 1.314e+000 | 1.433e+000 |
| 4 | 1.621e+000 | 1.315e+000 | 1.603e+000 | 1.446e+000 | 8.668e-001 | 1.336e+000 |
| 5 | 1.454e+000 | 8.668e-001 | 1.536e+000 | 1.312e+000 | 4.061e-001 | 1.167e+000 |
| 6 | 1.345e+000 | 4.496e-001 | 1.420e+000 | 1.114e+000 | 2.897e-001 | 9.808e-001 |
| 7 | 1.129e+000 | 1.605e-001 | 1.205e+000 | 9.250e-001 | 1.875e-001 | 7.974e-001 |
| 8 | 6.903e-001 | 3.335e-001 | 1.009e+000 | 7.405e-001 | 1.072e-001 | 6.113e-001 |
| 9 | 3.189e-001 | 8.653e-002 | 8.068e-001 | 5.605e-001 | 5.076e-002 | 4.407e-001 |
| 10 | 2.552e-001 | 9.040e-003 | 6.063e-001 | 3.855e-001 | 5.951e-003 | 2.679e-001 |
| 11 | 1.297e-001 | 3.398e-005 | 4.252e-001 | 2.016e-001 | 8.782e-006 | 1.058e-001 |
| 12 | 1.170e-002 | 6.333e-008 | 2.526e-001 | 3.005e-002 | 1.132e-007 | 1.903e-002 |
| 13 | 1.536e-004 | 2.641e-010 | 1.011e-001 | 5.290e-004 | 1.579e-009 | 5.277e-004 |
| 14 | 9.017e-008 | | 1.141e-002 | 1.367e-008 | | 8.758e-007 |
| 15 | 3.907e-010 | | 1.487e-004 | 1.135e-009 | | 3.347e-008 |
| 16 | | | 7.792e-008 | | | |

**7. Concluding remarks.** This paper presents a new formulation of iterative minimization to the saddle search problem. In this formulation, the problem is solved by iteratively solving a sequence of minimization subproblems. At each iteration, the rotation step of determining the softest eigenvector $v$ is followed by a nonlinear optimization for the subproblem to update the $x$ variable. We have proved the local quadratic convergence rate of the new scheme. This scheme is closely connected to the gentlest ascent dynamics (GAD) and other eigenvector-following algorithms such as the dimer method. However, our subproblem is not limited only on the direction $v$, but includes the information of the original energy function in all directions to update the $x$ variable in configuration space. The quadratic convergence rate theoretically established here is promising for further numerical improvement in practice and indicates that this would be the best rate for eigenvector-following–class algorithms. In a

forthcoming paper, we shall address the implementation of efficient algorithms based on this formulation. We are also interested in the saddle points of the free energy landscape in collective variables, where the free energy function $V$ is not known, but the force, the Hessian, and even the third order perturbation can be simultaneously computed from one single, but expensive, run of constrained molecular dynamics [18].

**Acknowledgment.** The authors would like to thank the anonymous referees for valuable suggestions on improvements in writing.

## REFERENCES

[1] P.-A. Absil, R. Mahony, and R. Sepulchre, *Optimization Algorithms on Matrix Manifolds*, Princeton University Press, Princeton, NJ, 2007.
[2] J. F. Bonnans and A. Shapiro, *Optimization problems with perturbations: A guided tour*, SIAM Rev., 40 (1998), pp. 228–264.
[3] C. A. Botsaris, *Constrained optimization along geodesics*, J. Math. Anal. Appl., 79 (1981), pp. 295–306.
[4] E. Cancès, F. Legoll, M.-C. Marinica, K. Minoukadeh, and F. Willaime, *Some improvements of the activation-relaxation technique method for finding transition pathways on potential energy surfaces*, J. Chem. Phys., 130 (2009), 114711.
[5] A. Cauchy, *Méthode générale pour la résolution des systèms d'equations simultanées*, C. R. Acad. Sci. Paris, 25 (1847), pp. 536–538.
[6] C. J. Cerjan and W. H. Miller, *On finding transition states*, J. Chem. Phys., 75 (1981), pp. 2800–2806.
[7] G. M. Crippen and H. A. Scheraga, *Minimization of polypeptide energy:* XI. *The method of gentlest ascent*, Arch. Biochem. Biophys., 144 (1971), pp. 462–466.
[8] Q. Du and L. Zhang, *A constrained string method and its numerical analysis*, Commun. Math. Sci., 7 (2009), pp. 1039–1051.
[9] W. E, W. Ren, and E. Vanden-Eijnden, *String method for the study of rare events*, Phys. Rev. B, 66 (2002), 052301.
[10] 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), 164103.
[11] W. E and X. Zhou, *The gentlest ascent dynamics*, Nonlinearity, 24 (2011), pp. 1831–1842.
[12] G. Henkelman, G. Jóhannesson, and H. Jónsson, *Methods for finding saddle points and minimum energy paths*, in Theoretical Methods in Condensed Phase Chemistry, Progr. Theor. Chem. Phys. 5, S. D. Schwartz, ed., Kluwer Academic Publishers, Dordrecht, The Netherlands, 2002, pp. 269–300.
[13] G. Henkelman and H. Jónsson, *A dimer method for finding saddle points on high dimensional potential surfaces using only first derivatives*, J. Chem. Phys., 111 (1999), pp. 7010–7022.
[14] A. Heyden, A. T. Bell, and F. J. Keil, *Efficient methods for finding transition states in chemical reactions: Comparison of improved dimer method and partitioned rational function optimization method*, J. Chem. Phys., 123 (2005), 224101.
[15] H. Jónsson, G. Mills, and K. W. Jacobsen, *Nudged elastic band method for finding minimum energy paths of transitions*, in Classical and Quantum Dynamics in Condensed Phase Simulations, Proceedings of the International School of Physics, LERICI, Villa Marigola, 1997, B. J. Berne, G. Ciccotti, and D. F. Coker, eds., World Scientific, Singapore, 1998, pp. 385–404.
[16] J. Kästner and P. Sherwood, *Superlinearly converging dimer method for transition state search*, J. Chem. Phys., 128 (2008), 014106.
[17] J. Leng, W. Gao, C. Shang, and Z.-P. Liu, *Efficient softest mode finding in transition states calculations*, J. Chem. Phys., 138 (2013), 094110.
[18] L. Maragliano, A. Fischer, E. Vanden-Eijnden, and G. Ciccotti, *String method in collective variables: Minimum free energy paths and isocommittor surfaces*, J. Chem. Phys., 125 (2006), 024106.
[19] P. Metzner, C. Schüttle, and E. Vanden-Eijnden, *Illustration of transition path theory on a collection of simple examples*, J. Chem. Phys., 125 (2006), 084110.
[20] N. Mousseau and G. T. Barkema, *Traveling through potential energy surfaces of disordered materials: The activation-relaxation technique*, Phys. Rev. E, 57 (1998), pp. 2419–2424.
[21] J. Nocedal, *Numerical Optimization*, Springer Ser. Oper. Res., Springer-Verlag, New York, 1999.

[22] S. Park, M. K. Sener, D. Lu, and K. Schulten, *Reaction paths based on mean first-passage times*, J. Chem. Phys., 119 (2003), pp. 1313–1319.
[23] W. Ren and E. Vanden-Eijnden, *A climbing string method for saddle point search*, J. Chem. Phys., 138 (2013), 134105.
[24] A. Samanta and W. E, *Atomistic simulations of rare events using gentlest ascent dynamics*, J. Chem. Phys., 136 (2012), 124104.
[25] D. J. Wales, *Energy Landscapes: Applications to Clusters, Biomolecules and Glasses*, Cambridge University Press, Cambridge, UK, 2003.
[26] J. Zhang and Q. Du, *Constrained shrinking dimer dynamics for saddle point search with constraints*, J. Comput. Phys., 231 (2012), pp. 4745–4758.
[27] J. Zhang and Q. Du, *Shrinking dimer dynamics and its applications to saddle point search*, SIAM J. Numer. Anal., 50 (2012), pp. 1899–1921.