CSIAM Trans. Appl. Math.
doi: 10.4208/csiam-am.SO-2022-0010

Vol. **4**, No. 1, pp. 157-176
March 2023

Mathematical and Numerical Analysis to Shrinking-Dimer Saddle Dynamics with Local Lipschitz Conditions

Lei Zhang[1], Pingwen Zhang[2] and Xiangcheng Zheng[3,*]

[1] *Beijing International Center for Mathematical Research, Center for Machine Learning Research, Center for Quantitative Biology, Peking University, Beijing 100871, China.*
[2] *School of Mathematical Sciences, Laboratory of Mathematics and Applied Mathematics, Peking University, Beijing 100871, China.*
[3] *School of Mathematics, Shandong University, Jinan 250100, China.*

Received 14 March 2022; Accepted 20 July 2022

Abstract. We present a mathematical and numerical investigation to the shrinking-dimer saddle dynamics for finding any-index saddle points in the solution landscape. Due to the dimer approximation of Hessian in saddle dynamics, the local Lipschitz assumptions and the strong nonlinearity for the saddle dynamics, it remains challenges for delicate analysis, such as the boundedness of the solutions and the dimer error. We address these issues to bound the solutions under proper relaxation parameters, based on which we prove the error estimates for numerical discretization to the shrinking-dimer saddle dynamics by matching the dimer length and the time step size. Furthermore, the Richardson extrapolation is employed to obtain a high-order approximation. The inherent reason of requiring the matching of the dimer length and the time step size lies in that the former serves a different mesh size from the later, and thus the proposed numerical method is close to a fully-discrete numerical scheme of some space-time PDE model with the Hessian in the saddle dynamics and its dimer approximation serving as a "spatial operator" and its discretization, respectively, which in turn indicates the PDE nature of the saddle dynamics.

AMS subject classifications: 37M05, 37N30, 65L20

Key words: Saddle dynamics, solution landscape, saddle points, local Lipschitz condition, error estimate, Richardson extrapolation.

*Corresponding author. *Email addresses:* zhangl@math.pku.edu.cn (L. Zhang), pzhang@pku.edu.cn (P. Zhang), zhengxch@outlook.com (X. Zheng)

1 Introduction

One of the major challenges in computational physical and chemistry is how to efficiently calculate saddle points on a complicated energy landscape. In comparison with finding local minima, the computation of saddle points is generally more difficult due to their unstable nature. Nevertheless, saddle points provide important information about the physical and chemical properties. For instance, the index-1 saddle point represents the transition states connecting two local minima according to the transition state theory [19, 35], and the index-2 saddle points are particularly interesting in chemical systems for providing valuable information on the trajectories of chemical reactions [15]. The applications of saddle points include nucleation in phase transformations [5, 32, 33], transition rates in chemical reactions and computational biology [11, 12, 21, 23, 25], etc.

The saddle points can be classified by the (Morse) index, which is characterized by the maximal dimension of a subspace on which the Hessian $H(x)$ is negative definite, according to the Morse theory [20]. Most existing searching algorithms focus on finding the index-1 saddle points, e.g. [1, 4, 6, 7, 9, 16, 17, 31]. However, the computation of high-index (index$>$1) saddle points receive less attention despite of the fact that the number of high-index saddles are much larger than the number of local minima and index-1 saddles on the complicated energy landscapes [2, 18].

The original saddle dynamics (SD) aims to find an index-k ($1\leq k\in\mathbb{N}$) saddle point of an energy function $E(x)$ [28]

$$\begin{cases} \dfrac{dx}{dt}=\beta\left(I-2\sum\limits_{j=1}^{k}v_jv_j^\top\right)F(x), \\ \dfrac{dv_i}{dt}=\gamma\left(I-v_iv_i^\top-2\sum\limits_{j=1}^{i-1}v_jv_j^\top\right)H(x)v_i, \quad 1\leq i\leq k. \end{cases} \tag{1.1}$$

Here the natural force $F:\mathbb{R}^N\to\mathbb{R}^N$ is generated from an energy $E(x)$ by $F(x)=-\nabla E(x)$, $H(x):=-\nabla^2E(x)$ corresponds to the Hessian of $E(x)$, β, $\gamma>0$ are relaxation parameters, x represents the position variable and direction variables $\{v_i\}_{i=1}^k$ form a basis for the unstable subspace of the Hessian at x.

Because the Hessians are often expensive to calculate and store, one can apply first derivatives to approximate the Hessians in Eq. (1.1) by using k dimers centered at x. To be specific, $H(x)v_i$ is approximated by

$$\hat{H}(x,v_i,l):=\frac{1}{2l}\big(F(x+lv_i)-F(x-lv_i)\big) \tag{1.2}$$

with the direction v_i and the dimer length $2l$ for some $l>0$.

Following the idea of the shrinking dimer dynamics [31, 34], we obtain the shrinking-dimer saddle dynamics (SSD) [28] as follows:

$$\begin{cases} \dfrac{dx}{dt} = \beta\left(I - 2\sum\limits_{j=1}^{k} v_j v_j^\top\right) F(x), \\ \dfrac{dv_i}{dt} = \gamma\left(I - v_i v_i^\top - 2\sum\limits_{j=1}^{i-1} v_j v_j^\top\right) \hat{H}(x, v_i, l), \quad 1 \le i \le k, \\ \dfrac{dl}{dt} = -l. \end{cases} \tag{1.3}$$

By using the SSD method as a key ingredient, the solution landscape can be constructed by connecting the high-index saddle points to low-index saddle points and local minima [26, 27]. The solution landscape serves as an efficient approach to provide a global structures of all stationary points of the model systems and has been widely applied in various fields [13, 14, 24–26, 29, 30].

Despite the growing applications of the SSD, the corresponding mathematical and numerical analysis are still far from well-developed. Most existing works only focus on the numerical analysis of the index-1 SD in recent years [8, 9, 17, 31], which corresponds to (1.1) with $k=1$, and the corresponding results for (shrinking-dimer) high-index SD are meager. In a very recent work [36], numerical discretization to SD (1.1) was analyzed, the proof of which depends heavily on the global Lipschitz assumptions of both $F(x)$ and $H(x)$. However, $F(x)$ and $H(x)$ generally have complex nonlinear forms that only admit local Lipschitz conditions. Furthermore, to avoid the direct calculation of Hessians, the SSD is usually used instead of the SD in practice. But, the dimer approximation of $H(x)$ in Eq. (1.3) introduces additional errors, which generate significant differences from the numerical analysis of the SD [36] and lead to the failure of the error estimates therein. Moreover, as the dimer length serves like a "step size" in the dimer approximation, it needs to be carefully chosen in order to match the time step size.

Motivated by these discussions, in this work we aim to prove the boundedness of the exact solutions and optimal-order error estimates of the numerical discretization to the SSD (1.3) with respect to the time step size. Due to the strong nonlinearity of the system and the local Lipschitz conditions, the boundedness of solutions is proved under some restrictions of the relaxation parameters (cf. (2.8)). Based on the proposed first-order scheme, the Richardson extrapolation is further developed to obtain a high-order approximation. As the dimer length serves as a different mesh size from the time step size, the proposed numerical method is close to a fully-discrete numerical scheme of some space-time PDE model with the Hessian and its dimer approximation serving as a "spatial operator" and its discretization, respectively, which in turn indicates the PDE nature of the saddle dynamics.

The rest of the paper is organized as follows. In Section 2 we estimate $x(t)$ and $\{v_i(t)\}_{i=1}^k$ in (1.3) under local Lipschitz conditions, which supports the subsequent numerical analysis. In Section 3 we present the numerical scheme of the SSD (1.3) and prove auxiliary estimates for the sake of the error estimates for the numerical discretization in Section 4. We also propose the Richardson extrapolation in this section to obtain a high-order approximation. In Section 5 we extend the developed techniques to numerically

analyze the generalized SSD for non-gradient systems. Numerical experiments are performed in Section 6 and we finally address a conclusion in the last section.

2 Estimate of solutions under local Lipschitz conditions

In this section we consider the SSD (1.3) on $[0,T]$, closed by the following initial conditions:

$$x(0)=x_0,\quad v_i(0)=v_{i,0}\quad \text{for}\quad 1\le i\le k,\quad v_{i,0}^\top v_{j,0}=\delta_{i,j},\quad l(0)=l_0. \tag{2.1}$$

We will show that, under the local Lipschitz conditions of $F(x)$ and $H(x)$, $x(t)$ and $\{v_i(t)\}_{i=1}^k$ with $0\le t\le T$ are bounded (under suitable relaxation parameters) such that in subsequent proofs, we could use the Lipschitz continuity of $F(x)$ and $H(x)$ with a fixed Lipschitz constant, just like imposing the global Lipschitz conditions as in [36]. Furthermore, the boundedness of $\{v_i(t)\}_{i=1}^k$ will be used in numerical analysis.

Let $\|\cdot\|$ be the standard l^2 norm of the matrix or the vector. For the sake of the analysis, we make the following assumption throughout the paper:

Assumption 2.1. $F(x)$ and $H(x)$ satisfy local Lipschitz conditions, that is, for any $r>0$ there exists a constant $L_r>0$ such that for $x_1,x_2\in B_r:=\{x\in\mathbb{R}^N:\|x\|\le r\}$

$$\|F(x_2)-F(x_1)\|\le L_r\|x_2-x_1\|,\quad \|H(x_2)-H(x_1)\|\le L_r\|x_2-x_1\|. \tag{2.2}$$

2.1 Properties of auxiliary functions

Based on Assumption 2.1, we derive some important properties for the following nonlinear functions:

$$\begin{aligned}
&X(x,v_1,\cdots,v_k):=\left(I-2\sum_{j=1}^k v_jv_j^\top\right)F(x),\\
&V_i(x,v_1,\cdots,v_k,l):=\left(I-v_iv_i^\top-2\sum_{j=1}^{i-1}v_jv_j^\top\right)\hat{H}(x,v_i,l),\quad 1\le i\le k,
\end{aligned}$$

which are indeed right-hand side terms of (1.3) without relaxation parameters, in the following theorem.

Theorem 2.1. *Under Assumption* 2.1, *for any fixed $r>0$ there exist positive constants $Q_0=Q_0(r)$ and $Q_2=Q_2(r)$ depending on r, L_r, k, l_0, F and H such that for $(x,v_1,\cdots,v_k)$, $(\bar{x},\bar{v}_1,\cdots,\bar{v}_k)\in\mathcal{B}_r$*

$$\begin{aligned}
&\|X(x,v_1,\cdots,v_k)-X(\bar{x},\bar{v}_1,\cdots,\bar{v}_k)\|\\
&\le Q_0(r)\|x-\bar{x},v_1-\bar{v}_1,\cdots,v_k-\bar{v}_k\|,
\end{aligned}\tag{2.3a}$$

$$\begin{aligned}
&\|V_i(x,v_1,\cdots,v_k,l)-V_i(\bar{x},\bar{v}_1,\cdots,\bar{v}_k,l)\|\\
&\le Q_2(r)\left(\|x-\bar{x},v_1-\bar{v}_1,\cdots,v_k-\bar{v}_k\|+l_0^2\right),\quad 1\le i\le k.
\end{aligned}\tag{2.3b}$$

Here the convex set $\mathcal{B}_r$ and the norm $\|x,v_1,\cdots,v_k\|$ are defined by

$$\mathcal{B}_r:=\left\{(x,v_1,\cdots,v_k):\|x,v_1,\cdots,v_k\|:=\left(\|x\|^2+\sum_{i=1}^k\|v_i\|^2\right)^{1/2}\leq r\right\}.$$

Remark 2.1. We write Q_0 and Q_2 as $Q_0(r)$ and $Q_2(r)$ in order to highlight their dependence on r. We neglect their dependence on k, l_0, F and H in the notations as these are fixed data throughout the paper.

Proof. Direct calculations show that for $(x,v_1,\cdots,v_k),(\bar{x},\bar{v}_1,\cdots,\bar{v}_k)\in\mathcal{B}_r$

$$\begin{aligned}
&\left\|X(x,v_1,\cdots,v_k)-X(\bar{x},\bar{v}_1,\cdots,\bar{v}_k)\right\|\\
\leq&\left\|\left(I-2\sum_{j=1}^k v_jv_j^\top\right)\big(F(x)-F(\bar{x})\big)\right\|+2\left\|\left(\sum_{j=1}^k\bar{v}_j\bar{v}_j^\top-\sum_{j=1}^k v_jv_j^\top\right)F(\bar{x})\right\|\\
\leq&(1+2kr^2)L_r\|x-\bar{x}\|+2\left\|\sum_{j=1}^k\bar{v}_j(\bar{v}_j^\top-v_j^\top)+(\bar{v}_j-v_j)v_j^\top\right\|\big(\|F(0)\|+L_r\|\bar{x}\|\big)\\
\leq&Q_0(r)\left\|x-\bar{x},v_1-\bar{v}_1,\cdots,v_k-\bar{v}_k\right\|,
\end{aligned}\tag{2.4}$$

where

$$Q_0(r):=\sqrt{k+1}\max\big\{(1+2kr^2)L_r,4r\big(\|F(0)\|+L_rr\big)\big\}.$$

To estimate $V_i(x,v_1,\cdots,v_k,l)-V_i(\bar{x},\bar{v}_1,\cdots,\bar{v}_k,l)$, we follow [9, Eq. (4)] to obtain

$$\hat{H}(x,v_i,l)=\frac{1}{2l}\big(F(x+lv_i)-F(x-lv_i)\big)=H(x)v_i+\mathcal{O}(l^2),\quad \mathcal{O}(l^2)\leq Q_1(r)l^2,\tag{2.5}$$

and consequently,

$$\left\|\hat{H}(x,v_i,l)-\hat{H}(\bar{x},\bar{v}_i,l)\right\|\leq\big(\|H(0)\|+rL_r\big)\big(\|v_i-\bar{v}_i\|+\|x-\bar{x}\|\big)+2Q_1(r)l_0^2.$$

We apply this relation and a similar derivation as (2.4) to obtain the Eq. (2.3b). Thus we complete the proof. □

2.2 Estimate of $(x,v_1,\cdots,v_k)$

Let $\varepsilon_0>0$ be a fixed constant. By Assumption 2.1, for a fixed r_0 satisfying

$$r_0\geq\|x_0,v_{1,0},\cdots,v_{k,0}\|+\varepsilon_0\tag{2.6}$$

there exists a constant $L_{r_0}>0$ such that (2.2) is satisfied. We then define $\tilde{X}(x,v_1,\cdots,v_k)$ and $\tilde{V}_i(x,v_1,\cdots,v_k,l)$ for $1\leq i\leq k$ such that

(i) $\tilde{X}=X,\quad \tilde{V}_i=V_i,\quad 1\leq i\leq k,\quad (x,v_1,\cdots,v_k)\in\mathcal{B}_{r_0}$;

(ii) $\tilde{X}$ and $\tilde{V}_i$ for $1\le i\le k$ satisfy the conditions (2.3) globally (i.e., for any choice of $(x,v_1,\cdots,v_k)$ and $(\bar{x},\bar{v}_1,\cdots,\bar{v}_k)$) with respect to the fixed constants $Q_0(r_0)$ and $Q_2(r_0)$.

Remark 2.2. A possible choice of $\tilde{X}$ is

$$\tilde{X}(x,v_1,\cdots,v_k)=\begin{cases} X(x,v_1,\cdots,v_k), & \text{if} \quad (x,v_1,\cdots,v_k)\in\mathcal{B}_{r_0},\\ X(\lambda x,\lambda v_1,\cdots,\lambda v_k), & \text{otherwise,}\end{cases}$$

where $\lambda:=\frac{r}{\|x,v_1,\cdots,v_k\|}$. The terms $\tilde{V}_1,\cdots,\tilde{V}_k$ could be similarly defined.

Consider the following modified SSD on $[0,T]$ with X and $\{V_i\}$ in (1.3) replaced by $\tilde{X}$ and $\{\tilde{V}_i\}$, respectively

$$\begin{cases}\dfrac{d\tilde{x}}{dt}=\beta\tilde{X}, & (2.7a)\\ \dfrac{d\tilde{v}_i}{dt}=\gamma\tilde{V}_i, \quad 1\le i\le k, & (2.7b)\\ \dfrac{dl}{dt}=-l, & (2.7c)\end{cases}$$

equipped with the initial conditions (2.1). We multiply $\tilde{x}$ on both sides of the Eq. (2.7a) and integrate the resulting equation from 0 to t to get

$$\begin{aligned}\|\tilde{x}(t)\|^2&\le\|x_0\|^2+2\beta\int_0^t\tilde{x}^\top\tilde{X}ds\\ &\le\|x_0\|^2+2\beta\int_0^t\|\tilde{x}(s)\|\Big[\big\|\tilde{X}(x_0,v_{1,0},\cdots,v_{k,0})\big\|\\ &\qquad+Q_0(r_0)\big(\|\tilde{x}(s),\tilde{v}_1(s),\cdots,\tilde{v}_k(s)\|+\|x_0,v_{1,0},\cdots,v_{k,0}\|\big)\Big]ds\\ &\le\|x_0\|^2+\big(\beta+Q_0(r_0)\beta\big)\int_0^t\big\|\tilde{x}(s)\big\|^2ds+\beta T\big(\|X(x_0,v_{1,0},\cdots,v_{k,0})\|+Q_0(r_0)r_0\big)^2\\ &\quad+\beta Q_0(r_0)\int_0^t\big\|\tilde{x}(s),\tilde{v}_1(s),\cdots,\tilde{v}_k(s)\big\|^2ds,\end{aligned}$$

where we used $(x_0,v_{1,0},\cdots,v_{k,0})\in\mathcal{B}_{r_0}$ in this derivation. Similarly, we bound $\tilde{v}_i$ in (2.7) by

$$\begin{aligned}\big\|\tilde{v}_i(t)\big\|^2&\le\|v_{i,0}\|^2+\big(\gamma+Q_2(r_0)\gamma\big)\int_0^t\|\tilde{v}_i(s)\|^2ds\\ &\quad+\gamma T\big(\|V_i(x_0,v_{1,0},\cdots,v_{k,0})\|+Q_2(r_0)r_0+Q_2(r_0)l_0^2\big)^2\\ &\quad+\gamma Q_2(r_0)\int_0^t\big\|\tilde{x}(s),\tilde{v}_1(s),\cdots,\tilde{v}_k(s)\big\|^2ds.\end{aligned}$$

Furthermore, we apply (2.5) to obtain

$$\big\|X(x_0,v_{1,0},\cdots,v_{k,0})\big\|\le\left(1+2k\sum_{j=1}^k\|v_{j,0}\|^2\right)\|F(x_0)\|=:Q_3,$$

$$\begin{aligned}\left\|V_i(x_0,v_{1,0},\cdots,v_{k,0})\right\| \leq &\left(1+\|v_{i,0}\|^2+2k\sum_{j=1}^{k}\|v_{j,0}\|^2\right)\\ &\times\left(\|H(x_0)\|\|v_{i,0}\|+Q_1(r_0)l_0^2\right)=:Q_4(r_0).\end{aligned}$$

We incorporate the above equations to obtain

$$\begin{aligned}&\|\tilde{x}(t),\tilde{v}_1(t),\cdots,\tilde{v}_k(t)\|^2\\ &\leq\|x_0,v_{1,0},\cdots,v_{k,0}\|^2+Q_5(r_0)+Q_6(r_0)\int_0^t\|\tilde{x}(s),\tilde{v}_1(s),\cdots,\tilde{v}_k(s)\|^2ds,\end{aligned}$$

where

$$\begin{aligned}Q_5(r_0)&:=\beta T\left(Q_3+Q_0(r_0)r_0\right)^2+k\gamma T\left(Q_4(r_0)+Q_2(r_0)r_0+Q_2(r_0)l_0^2\right)^2,\\ Q_6(r_0)&:=\max\left\{\beta\left(1+Q_0(r_0)\right),\gamma\left(1+Q_2(r_0)\right)\right\}+\beta Q_0(r_0)+k\gamma Q_2(r_0).\end{aligned}$$

Then an application of the Gronwall's inequality yields

$$\left\|\tilde{x}(t),\tilde{v}_1(t),\cdots,\tilde{v}_k(t)\right\|\leq\left(\|x_0,v_{1,0},\cdots,v_{k,0}\|^2+Q_5(r_0)\right)^{1/2}e^{Q_6(r_0)T/2}=:S(\beta,\gamma)$$

for $0\leq t\leq T$. As $S(\beta,\gamma)$ is an increasing function with respect to both β and γ, where $\beta,\gamma\geq 0$ and attains its minimum $\|x_0,v_{1,0},\cdots,v_{k,0}\|$ at $\beta=\gamma=0$, we could select β and γ such that

$$S(\beta,\gamma)\leq\|x_0,v_{1,0},\cdots,v_{k,0}\|+\varepsilon_0, \tag{2.8}$$

which implies

$$\|\tilde{x}(t),\tilde{v}_1(t),\cdots,\tilde{v}_k(t)\|\leq\|x_0,v_{1,0},\cdots,v_{k,0}\|+\varepsilon_0,\quad 0\leq t\leq T. \tag{2.9}$$

Recalling that $r_0\geq\|x_0,v_{1,0},\cdots,v_{k,0}\|+\varepsilon_0$, we base on (i) to conclude that the modified SSD (2.7) is indeed equivalent to the original SSD (1.3) for (β,γ) satisfying (2.8). We summarize the findings in the following theorem.

Theorem 2.2. *Suppose Assumption* 2.1 *holds and* (β,γ) *satisfy* (2.8)*, then* $(x,v_1,\cdots,v_k)$ *in the SSD* (1.3) *are bounded as* (2.9) *and thus we could always apply the Lipschitz conditions of* $F(x)$ *and* $H(x)$ *in subsequent proofs with a fixed Lipschitz constant* L_{r_0} *corresponding to the* r_0 *given by* (2.6)*.*

3 Discrete SSD and auxiliary results

In this section we present the numerical scheme to (1.3) and prove auxiliary lemmas to be used in the error estimates.

3.1 Numerical scheme

Let $0=t_0<t_1<\cdots<t_K=T$ be a uniform temporal partition of $[0,T]$ with the time step size $\tau:=T/K$ for some $0<K\in\mathbb{N}$. We approximate the first-order derivative by the Euler scheme at t_n as follows:

$$\frac{dg(t_n)}{dt}=\frac{1}{\tau}\big(g(t_n)-g(t_{n-1})\big)+R_n^g,$$

where g refers to x or v_i, and we suppose the truncation error satisfies $\|R_n^g\|=\mathcal{O}(\tau)$. Invoking this discretization in (1.3) yields the following reference equations for the dynamics (1.3):

$$\begin{cases} x(t_n)=x(t_{n-1})+\tau\beta\Big(I-2\sum_{j=1}^{k}v_j(t_{n-1})v_j^\top(t_{n-1})\Big)F\big(x(t_{n-1})\big)+\tau R_n^x, & (3.1a)\\ v_i(t_n)=v_i(t_{n-1})+\tau\gamma\Big(I-v_i(t_{n-1})v_i^\top(t_{n-1})-2\sum_{j=1}^{i-1}v_j(t_{n-1})v_j^\top(t_{n-1})\Big)\\ \qquad\times\hat{H}\big(x(t_{n-1}),v_i(t_{n-1}),l(t_{n-1})\big)+\tau R_n^{v_i},\quad 1\le i\le k, & (3.1b)\\ l(t_n)=e^{-t_n}l_0, & (3.1c)\end{cases}$$

where we analytically solved the equation of l without approximation. In the rest of the paper we denote

$$l_n=l(t_n)=e^{-t_n}l_0$$

for simplicity and we then drop the truncation errors in the reference equations to obtain the explicit scheme of (1.3)

$$\begin{cases} x_n=x_{n-1}+\tau\beta\Big(I-2\sum_{j=1}^{k}v_{j,n-1}v_{j,n-1}^\top\Big)F(x_{n-1}), & (3.2a)\\ \tilde{v}_{i,n}=v_{i,n-1}+\tau\gamma\Big(I-v_{i,n-1}v_{i,n-1}^\top-2\sum_{j=1}^{i-1}v_{j,n-1}v_{j,n-1}^\top\Big)\\ \qquad\times\hat{H}(x_{n-1},v_{i,n-1},l_{n-1}),\quad 1\le i\le k, & (3.2b)\\ \{v_{i,n}\}_{i=1}^k=\mathrm{GS}\big(\{\tilde{v}_{i,n}\}_{i=1}^k\big) & (3.2c)\end{cases}$$

for $1\le n\le K$, equipped with the initial conditions (2.1). Here the notation $\mathrm{GS}(\cdot)$ refers to the Gram-Schmidt orthonormalization procedure, the purpose of which is to preserve the orthonormal property of the vectors [27,28]. Due to the orthonormalization, $\|v_{i,n}\|=1$ for all possible i and n, and by a discrete analogue of the derivations in Section 2, we could obtain the estimate of x_n. To be specific, let $r_0\ge\|x_0\|+\varepsilon_0$ be a fixed constant for some $\varepsilon_0>0$. Then we consider the following auxiliary problem:

$$\tilde{x}_n=\tilde{x}_{n-1}+\tau\beta\Big(I-2\sum_{j=1}^{k}v_{j,n-1}v_{j,n-1}^\top\Big)\tilde{F}(\tilde{x}_{n-1}) \tag{3.3}$$

for $1\leq n\leq K$ with $\tilde{x}_0=x_0$. Here $\tilde{F}(\cdot)$ is defined as

$$\tilde{F}(x)=\begin{cases} F(x), & \|x\|\leq r_0, \\ F\left(\dfrac{r_0}{\|x\|}x\right), & \text{otherwise}, \end{cases}$$

such that $\tilde{F}$ satisfies the global Lipschitz condition with the Lipschitz constant L_{r_0}. Then we apply the norm-preserving property of the Householder matrix

$$I-2\sum_{j=1}^{k}v_{j,n-1}v_{j,n-1}^{\top}$$

on (3.3) to obtain

$$\|\tilde{x}_n\|\leq\|\tilde{x}_{n-1}\|+\tau\beta\|\tilde{F}(\tilde{x}_{n-1})\|.$$

We then apply the global Lipschitz condition of $\tilde{F}$ to get

$$\begin{aligned}\|\tilde{x}_n\|&\leq\|\tilde{x}_{n-1}\|+\tau\beta\big(\|\tilde{F}(0)\|+L_{r_0}\|\tilde{x}_{n-1}\|\big)\\&=\|\tilde{x}_{n-1}\|+\tau\beta\big(\|F(0)\|+L_{r_0}\|\tilde{x}_{n-1}\|\big).\end{aligned}$$

Adding this equation for $1\leq n\leq m$ for some $m\leq K$ yields

$$\|\tilde{x}_m\|\leq\|x_0\|+\beta T\|F(0)\|+\tau\beta L_{r_0}\sum_{n=1}^{m}\|\tilde{x}_{n-1}\|.$$

Then an application of the discrete Gronwall inequality leads to

$$\|\tilde{x}_m\|\leq\big(\|x_0\|+\beta T\|F(0)\|\big)e^{\beta L_{r_0}T}.$$

Consequently, if β satisfies

$$\big(\|x_0\|+\beta T\|F(0)\|\big)e^{\beta L_{r_0}T}\leq\|x_0\|+\varepsilon_0, \tag{3.4}$$

$\tilde{x}_n$ is bounded as

$$\|\tilde{x}_n\|\leq\|x_0\|+\varepsilon_0,\quad 0\leq n\leq K, \tag{3.5}$$

and thus the Eq. (3.3) is equivalent to the Eq. (3.2a). This implies that x_n is bounded as (3.5) and we could always apply the Lipschitz conditions of $F(x)$ and $H(x)$ in subsequent proofs with a fixed Lipschitz constant L_{r_0}.

In the rest of the paper we use Q to denote a generic positive constant that may assume different values at different occurrences.

3.2 Auxiliary estimates

We prove several auxiliary estimates to support the error estimates.

Lemma 3.1. *Suppose* (2.8), (3.4) *and Assumption* 2.1 *hold and* $l_0^2=\mathcal{O}(\tau)$, *then the following estimates hold for* $1\le n\le K$:

$$\left|\tilde{v}_{m,n}^\top \tilde{v}_{i,n}\right|\le Q\tau^2,\quad 1\le m<i\le k.$$

Here the positive constant Q *is independent from* n, K, *and* τ.

Remark 3.1. Note that the initial value l_0 of the dimer length l is chosen in the magnitude of $\sqrt{\tau}$, which is key in preserving the first-order accuracy of the scheme (3.2) as we will see later. The inherent reason is that the numerical method (3.2) is close to a fully-discrete numerical scheme of some space-time PDE model with the Hessian and its dimer approximation serving as a "spatial operator" and its discretization, respectively, and the dimer length serves like a "spatial mesh size", which should match the time-stepping size τ to keep the $\mathcal{O}(\tau)$ accuracy of the numerical method.

Proof. For $1\le m<i\le k$ we apply (2.5) and the symmetry of $H(x)$ to obtain

$$\left|v_{i,n-1}^\top \hat{H}(x_{n-1},v_{m,n-1},l_{n-1})-v_{m,n-1}^\top \hat{H}(x_{n-1},v_{i,n-1},l_{n-1})\right|=\left|\mathcal{O}(l_{n-1}^2)\right|\le Q\tau. \tag{3.6}$$

By Assumption 2.1 we conclude that $H(x_n)$, $0\le n\le K$ is bounded, which, together with (2.5), implies that $\hat{H}(x_n,v_{j,n},l_n)$ for $0\le n\le K$, $1\le j\le k$ are bounded. We invoke these boundedness and (3.6) to the right-hand side of the following equation:

$$\tilde{v}_{m,n}^\top \tilde{v}_{i,n}=\tau\gamma\left(v_{i,n-1}^\top \hat{H}(x_{n-1},v_{m,n-1},l_{n-1})-v_{m,n-1}^\top \hat{H}(x_{n-1},v_{i,n-1},l_{n-1})\right)+\mathcal{O}(\tau^2) \tag{3.7}$$

to complete the proof. □

Lemma 3.2. *Under* (2.8), (3.4) *and Assumption* 2.1, *the following estimates hold for* $1\le n\le K$:

$$\left|\|\tilde{v}_{i,n}\|^2-1\right|\le Q\tau^2,\quad 1\le i\le k.$$

Here the positive constant Q *is independent from* n, K *and* τ.

Proof. By the boundedness of $\hat{H}$, cf. the proof of Lemma 3.1, we obtain from the Eq. (3.2b) that

$$\|\tilde{v}_{i,n}-v_{i,n-1}\|=\tau\gamma\left\|\left(I-v_{i,n-1}v_{i,n-1}^\top-2\sum_{j=1}^{i-1}v_{j,n-1}v_{j,n-1}^\top\right)\hat{H}(x_{n-1},v_{i,n-1},l_{n-1})\right\|\le Q\tau. \tag{3.8}$$

We multiply $v_{i,n-1}^\top$ on both sides of the Eq. (3.2b) and use the orthonormal property of $\{v_{i,n-1}\}_{i=1}^k$ to obtain for $1\le i\le k$

$$\begin{aligned}v_{i,n-1}^\top \tilde{v}_{i,n}=v_{i,n-1}^\top v_{i,n-1}+\tau\gamma\Bigg(&v_{i,n-1}^\top-v_{i,n-1}^\top v_{i,n-1}v_{i,n-1}^\top-2\sum_{j=1}^{i-1}v_{i,n-1}^\top v_{j,n-1}v_{j,n-1}^\top\Bigg)\\&\times\hat{H}(x_{n-1},v_{i,n-1},l_{n-1})=1.\end{aligned} \tag{3.9}$$

We then multiply $\tilde{v}_{i,n}^\top$ on both sides of the Eq. (3.2b) and apply (3.9) and the orthogonality of $\{v_{i,n-1}\}_{i=1}^k$ to obtain

$$\begin{aligned}\tilde{v}_{i,n}^\top\tilde{v}_{i,n}&=\tilde{v}_{i,n}^\top v_{i,n-1}+\tau\gamma\left(\tilde{v}_{i,n}^\top-\tilde{v}_{i,n}^\top v_{i,n-1}v_{i,n-1}^\top-2\sum_{j=1}^{i-1}\tilde{v}_{i,n}^\top v_{j,n-1}v_{j,n-1}^\top\right)\hat{H}(x_{n-1},v_{i,n-1},l_{n-1})\\&=1+\tau\gamma\left(\tilde{v}_{i,n}^\top-v_{i,n-1}^\top-2\sum_{j=1}^{i-1}(\tilde{v}_{i,n}-v_{i,n-1})^\top v_{j,n-1}v_{j,n-1}^\top\right)\hat{H}(x_{n-1},v_{i,n-1},l_{n-1}).\end{aligned}\tag{3.10}$$

Invoking (3.8) in (3.10) leads to

$$\begin{aligned}\left|\|\tilde{v}_{i,n}\|^2-1\right|&\le\tau\gamma\left(\|\tilde{v}_{i,n}-v_{i,n-1}\|+2\sum_{j=1}^{i-1}\|\tilde{v}_{i,n}-v_{i,n-1}\|\right)\left\|\hat{H}(x_{n-1},v_{i,n-1},l_{n-1})\right\|\\&\le Q\tau^2,\quad 1\le i\le k,\quad 1\le n\le K,\end{aligned}$$

which completes the proof. □

Lemma 3.3. *Suppose* (2.8), (3.4) *and Assumption* 2.1 *hold and* $l_0^2=\mathcal{O}(\tau)$*, the following estimate holds for* $1\le n\le K$ *and* τ *sufficiently small:*

$$\|v_{i,n}-\tilde{v}_{i,n}\|\le Q\tau^2,\quad 1\le i\le k.$$

Here the positive constant Q *is independent from* n*,* K *and* τ*.*

Proof. The proof could be performed following that of [36, Lemma 4.2] and is thus omitted. □

4 Error estimate and accuracy improvement

In this section we prove error estimates for the numerical discretization (3.2) to the SSD (1.3). Based on the analyzed first-order scheme (3.2), we then employ the Richardson extrapolation to obtain a second-order approximation.

4.1 Error estimate of (3.2)

We analyze the scheme (3.2) in the following theorem.

Theorem 4.1. *Suppose* (2.8), (3.4) *and Assumption* 2.1 *hold and* $l_0^2=\mathcal{O}(\tau)$*. Then the following estimate holds for* τ *sufficiently small:*

$$\max_{1\le n\le K}\left(\|x(t_n)-x_n\|+\sum_{i=1}^k\|v_i(t_n)-v_{i,n}\|\right)\le Q\tau.$$

Here Q *is independent from* τ*,* n *and* K*.*

Proof. Let

$$e_n^x := x(t_n) - x_n, \quad e_n^{v_i} := v_i(t_n) - v_{i,n} \tag{4.1}$$

and we subtract the Eq. (3.1b) from that of (3.2) and apply the splitting

$$v_i(t_n) - \tilde{v}_{i,n} = e_n^{v_i} + (v_{i,n} - \tilde{v}_{i,n})$$

to obtain

$$\begin{aligned} e_n^{v_i} = e_{n-1}^{v_i} &+ \tau\gamma\big(\hat{H}\big(x(t_{n-1}), v_i(t_{n-1}), l(t_{n-1})\big) - \hat{H}(x_{n-1}, v_{i,n-1}, l_{n-1})\big) \\ &- \tau\gamma\Big[v_i(t_{n-1})v_i(t_{n-1})^\top \hat{H}\big(x(t_{n-1}), v_i(t_{n-1}), l(t_{n-1})\big) \\ &\qquad - v_{i,n-1}v_{i,n-1}^\top \hat{H}(x_{n-1}, v_{i,n-1}, l_{n-1})\Big] \\ &- 2\tau\gamma\sum_{j=1}^{i-1}\Big[v_j(t_{n-1})v_j(t_{n-1})^\top \hat{H}\big(x(t_{n-1}), v_i(t_{n-1}), l(t_{n-1})\big) \\ &\qquad - v_{j,n-1}v_{j,n-1}^\top \hat{H}(x_{n-1}, v_{i,n-1}, l_{n-1})\Big] - (v_{i,n} - \tilde{v}_{i,n}) + \tau R_n^{v_i}. \end{aligned} \tag{4.2}$$

By (2.5) we bound the first difference on the right-hand side of (4.2)

$$\begin{aligned} &\big\|\hat{H}\big(x(t_{n-1}), v_i(t_{n-1}), l(t_{n-1})\big) - \hat{H}(x_{n-1}, v_{i,n-1}, l_{n-1})\big\| \\ &= \big\|H\big(x(t_{n-1})\big)v_i(t_{n-1}) - H(x_{n-1})v_{i,n-1} + \mathcal{O}(\tau)\big\| \\ &= \big\|H\big(x(t_{n-1})\big)\big(v_i(t_{n-1}) - v_{i,n-1}\big) + \big(H\big(x(t_{n-1})\big) - H(x_{n-1})\big)v_{i,n-1} + \mathcal{O}(\tau)\big\| \\ &\le Q\|e_{n-1}^{v_i}\| + Q\|e_{n-1}^x\| + Q\tau. \end{aligned} \tag{4.3}$$

To generate errors from other differences on the right-hand side of (4.2), we should introduce several intermediate terms to split them. For instance, the second difference on the right-hand side of (4.2) could be split as

$$\begin{aligned} &\big\|v_i(t_{n-1})v_i(t_{n-1})^\top \hat{H}\big(x(t_{n-1}), v_i(t_{n-1}), l(t_{n-1})\big) - v_{i,n-1}v_{i,n-1}^\top \hat{H}(x_{n-1}, v_{i,n-1}, l_{n-1})\big\| \\ &= \big\|e_{n-1}^{v_i}v_i(t_{n-1})^\top \hat{H}\big(x(t_{n-1}), v_i(t_{n-1}), l(t_{n-1})\big) + v_{i,n-1}(e_{n-1}^{v_i})^\top \hat{H}(x_{n-1}, v_{i,n-1}, l_{n-1}) \\ &\quad + v_{i,n-1}v_{i,n-1}^\top\big(\hat{H}\big(x(t_{n-1}), v_i(t_{n-1}), l(t_{n-1})\big) - \hat{H}(x_{n-1}, v_{i,n-1}, l_{n-1})\big)\big\| \\ &\le Q\big(\|e_{n-1}^{v_i}\| + \|e_{n-1}^x\| + \tau\big), \end{aligned}$$

where we used (4.3) and the boundedness of $\hat{H}$ in the last estimate. The other differences on the right-hand side of (4.2) could be estimated similarly. We incorporate these estimates in (4.2) and apply Lemma 3.3 to obtain

$$\|e_n^{v_i}\| \le \|e_{n-1}^{v_i}\| + Q\tau\big(\|e_{n-1}^x\| + \|e_{n-1}^v\|\big) + Q\tau^2, \tag{4.4}$$

where

$$\|e_n^v\| := \sum_{j=1}^k \big\|e_n^{v_j}\big\|. \tag{4.5}$$

We then subtract the Eq. (3.1a) from that of (3.2) to obtain

$$\begin{aligned}
e_n^x &= e_{n-1}^x + \tau\beta\big(F\big(x(t_{n-1})\big) - F(x_{n-1})\big) \\
&\quad -2\tau\beta\sum_{j=1}^{k}\Big[v_j(t_{n-1})v_j(t_{n-1})^\top F\big(x(t_{n-1})\big) - v_{j,n-1}v_{j,n-1}^\top F(x_{n-1})\Big] + \tau R_n^x \\
&= e_{n-1}^x + \tau\beta(F(x(t_{n-1})) - F(x_{n-1})) \\
&\quad -2\tau\beta\sum_{j=1}^{k}\Big[e_{n-1}^{v_j}v_j(t_{n-1})^\top F(x(t_{n-1})) + v_{j,n-1}(e_{n-1}^{v_j})^\top F(x(t_{n-1})) \\
&\qquad\qquad + v_{j,n-1}v_{j,n-1}^\top\big(F(x(t_{n-1})) - F(x_{n-1})\big)\Big] + \tau R_n^x.
\end{aligned}$$

Similar to the above derivations, we apply Assumption 2.1, the boundedness of $\|F\|$ and $\|R_n^x\| = \mathcal{O}(\tau)$ to find

$$\|e_n^x\| \le \|e_{n-1}^x\| + Q\tau\|e_{n-1}^x\| + Q\tau\|e_{n-1}^v\| + Q\tau^2.$$

Adding this equation from $n=1$ to m yields

$$\|e_m^x\| \le Q\tau\sum_{n=1}^{m}\|e_{n-1}^x\| + Q\tau\sum_{n=1}^{m}\|e_{n-1}^v\| + Q\tau.$$

Then an application of the discrete Gronwall inequality leads to

$$\|e_n^x\| \le Q\tau\sum_{m=1}^{n-1}\|e_m^v\| + Q\tau, \quad 1 \le n \le K. \tag{4.6}$$

We invoke this equation in (4.4) to obtain

$$\|e_n^{v_i}\| \le \|e_{n-1}^{v_i}\| + Q\tau\|e_{n-1}^v\| + Q\tau^2\sum_{m=1}^{n-1}\|e_m^v\| + Q\tau^2.$$

We then sum up this equation for $1 \le i \le k$ to get

$$\|e_n^v\| \le \|e_{n-1}^v\| + Q\tau\|e_{n-1}^v\| + Q\tau^2\sum_{m=1}^{n-1}\|e_m^v\| + Q\tau^2.$$

Adding this equation from $n=1$ to n_* leads to

$$\|e_{n^*}^v\| \le Q\tau\sum_{n=1}^{n_*}\|e_{n-1}^v\| + Q\tau^2\sum_{n=1}^{n_*}\sum_{m=1}^{n-1}\|e_m^v\| + Q\tau \le Q\tau\sum_{n=1}^{n_*}\|e_{n-1}^v\| + Q\tau.$$

Then an application of the discrete Gronwall inequality again yields

$$\|e_n^v\| \le Q\tau, \quad 1 \le n \le K,$$

and we combine this with (4.6) to obtain the estimate of $\|e_n^x\|$, which completes the proof of this theorem. □

4.2 A second-order accuracy technique

In Section 4.1, we show that the scheme (3.2) has the first-order accuracy. A useful approach to get the high-order approximations from the low-order ones is the Richardson extrapolation (see e.g., [3, 22]), which is a smart combination of numerical solutions of low-order schemes under different partitions to reach high-order accuracy. A typical and simple example is the second-order Richardson extrapolation. Let

$$\{x_n, v_{1,n}, \cdots, v_{k,n}\}_{n=0}^{K}, \quad \{\bar{x}_m, \bar{v}_{1,m}, \cdots, \bar{v}_{k,m}\}_{m=0}^{2K}$$

be numerical solutions of the first-order scheme (3.2) with the mesh numbers K and $2K$, respectively. Then the Richardson extrapolation yields the approximation solution

$$\{x_n^R, v_{1,n}^R, \cdots, v_{k,n}^R\}_{n=0}^{K}$$

of second-order accuracy on the coarse mesh defined as

$$x_n^R = 2\bar{x}_{2n} - x_n, \quad v_{i,n}^R = 2\bar{v}_{i,2n} - v_{i,n}, \quad 1 \le i \le k, \quad 0 \le n \le K.$$

The analysis of the second-order accuracy is standard and we refer [22, Section 9.6] for details.

5 Generalized SSD of non-gradient systems

In many autonomous dynamical systems there exists no energy $E(x)$ such that $F(x) = -\nabla E(x)$, that is, these systems are non-gradient dynamics. In this case, the following SD is developed in [27] via the Jacobian $J(x) = \nabla F(x)$ to search for the saddle points of non-gradient systems:

$$\begin{cases} \dfrac{dx}{dt} = \left(I - 2\sum\limits_{j=1}^{k} v_j v_j^\top\right) F(x), \\ \dfrac{dv_i}{dt} = \left(I - v_i v_i^\top\right) \nabla F \cdot v_i - \sum\limits_{j=1}^{i-1} v_j v_j^\top \left(J(x) + J(x)^\top\right) v_i, \quad 1 \le i \le k. \end{cases} \tag{5.1}$$

If we again employ the dimer method as (1.3) to approximate the multiplication of the Jacobian and the vector for efficient implementation, then the following generalized SSD could be derived from (5.1):

$$\begin{cases} \dfrac{dx}{dt} = \left(I - 2\sum\limits_{j=1}^{k} v_j v_j^\top\right) F(x), \\ \dfrac{dv_i}{dt} = \left(I - v_i v_i^\top\right) \hat{H}(x, v_i, l) \\ \qquad\quad - \sum\limits_{j=1}^{i-1} v_j \left(v_j^\top \hat{H}(x, v_i, l) + v_i^\top \hat{H}(x, v_j, l)\right), \quad 1 \le i \le k, \\ \dfrac{dl}{dt} = -l. \end{cases} \tag{5.2}$$

Compared with the SSD (1.3), a symmetrization $v_j^\top \hat{H}(x,v_i,l)+v_i^\top \hat{H}(x,v_j,l)$ is used to replace $2v_j^\top \hat{H}(x,v_i,l)$ in the dynamics of v_i in response to the asymmetry of ∇F. Similar to Section 3, the corresponding numerical scheme to (5.2) reads

$$\begin{cases} x_n = x_{n-1}+\tau\left(I-2\sum_{j=1}^{k} v_{j,n-1}v_{j,n-1}^\top\right)F(x_{n-1}), \\ \tilde{v}_{i,n} = v_{i,n-1}+\tau\big(I-v_{i,n-1}v_{i,n-1}^\top\big)\hat{H}(x_{n-1},v_{i,n-1},l_{n-1}) \\ \qquad -\tau\sum_{j=1}^{i-1} v_{j,n-1}\big(v_{j,n-1}^\top \hat{H}(x_{n-1},v_{i,n-1},l_{n-1}) \\ \qquad\qquad +v_{i,n-1}^\top \hat{H}(x_{n-1},v_{j,n-1},l_{n-1})\big), \qquad 1\le i\le k, \\ \{v_{i,n}\}_{i=1}^k = \mathrm{GS}\big(\{\tilde{v}_{i,n}\}_{i=1}^k\big). \end{cases} \tag{5.3}$$

We may follow the preceding proofs to analyze the scheme (5.3). However, a key difference that may lead to the failure of recycling the developed ideas and techniques lies in the estimate (3.7) of $\tilde{v}_{m,n}^\top \tilde{v}_{i,n}$ for $1\le m<i\le k$, which is delicate as we require $\mathcal{O}(\tau^2)$ accuracy. Therefore, we reestimate this term for the scheme (5.3) as follows:

$$\begin{aligned} \tilde{v}_{m,n}^\top \tilde{v}_{i,n} = \tau\Big[& v_{m,n-1}^\top \hat{H}(x_{n-1},v_{i,n-1},l_{n-1}) \\ & -\big(v_{m,n-1}^\top \hat{H}(x_{n-1},v_{i,n-1},l_{n-1})+v_{i,n-1}^\top \hat{H}(x_{n-1},v_{m,n-1},l_{n-1})\big) \\ & +\big(\hat{H}(x_{n-1},v_{m,n-1},l_{n-1})\big)^\top v_{i,n-1}\Big]+\mathcal{O}(\tau^2)=\mathcal{O}(\tau^2), \end{aligned}$$

where we used the observation that the content in $[\cdots]$ in the last-but-one equality is exactly 0 by virtue of the symmetrization. The other proofs could be performed in parallel to prove first-order accuracy for all variables in the numerical scheme (5.3) of the generalized SSD (5.2) for non-gradient systems, and the Richardson extrapolation proposed in Section 4.2 could also be employed to obtain the approximate solutions of second-order accuracy.

6 Numerical experiments

In this section, we carry out numerical experiments to substantiate the accuracy of the numerical schemes (3.2) and (5.3). For applications of these schemes in practical problems, we refer [27, 28] for various physical examples and detailed discussions. As the exact solutions to the dynamics are not available, numerical solutions computed under $\tau=2^{-13}$ serve as the reference solutions. In the following examples, we set $\beta=\gamma=T=1$ for simplicity and denote the convergence rate by CR. The errors $\|e_n^x\|$ and $\|e_n^v\|$ measured in the experiments are defined in (4.1), (4.5), and we further define the norms

$$\big\|e_n^{R,x}\big\| := \big\|x(t_n)-x_n^R\big\|, \quad \big\|e_n^{R,v}\big\| := \sum_{j=1}^{k}\big\|v_j(t_n)-v_{j,n}^R\big\|$$

for the errors of the Richardson extrapolation. In all experiments, l_0 is chosen as $\sqrt{\tau}$.

6.1 First-order scheme for gradient system

We consider the SSD (1.3) for the stingray function $E(x_1,x_2)=x_1^2+(x_1-1)x_2^2$ [10] and compute its index-1 and index-2 saddle points via scheme (3.2) with the initial conditions $x_0=(1,1)^\top$, $v_0=(0,1)^\top$ and $x_0=(1,1)^\top$, $v_{1,0}=(0,1)^\top$, $v_{2,0}=(1,0)^\top$, respectively. Numerical results are presented in Tables 1-2, which demonstrate the first-order accuracy of the numerical scheme (3.2) as proved in Section 4.

Table 1: Convergence of (3.2) for finding an index-1 saddle point in Example 6.1.

$1/\tau$	$\max_n \|e_n^x\|$	CR	$\max_n \|e_n^v\|$	CR
2^5	2.60E-02		1.91E-02	
2^6	1.23E-02	1.08	9.22E-03	1.05
2^7	5.98E-03	1.05	4.51E-03	1.03
2^8	2.91E-03	1.04	2.20E-03	1.03

Table 2: Convergence of (3.2) for finding an index-2 saddle point in Example 6.1.

$1/\tau$	$\max_n \|e_n^x\|$	CR	$\max_n \|e_n^v\|$	CR
2^5	1.50E-02		3.90E-02	
2^6	7.41E-03	1.02	1.90E-02	1.04
2^7	3.66E-03	1.02	9.30E-03	1.03
2^8	1.79E-03	1.03	4.55E-03	1.03

6.2 First-order scheme for non-gradient system

We consider the following (non-gradient) dynamical system:

$$\frac{dx}{dt}=\begin{bmatrix} 1 & 0.5 & 0 \\ -0.5 & 1 & -0.3 \\ 0 & -0.2 & 1 \end{bmatrix} x+\begin{bmatrix} (1+(x_1-1)^2)^{-1} \\ (1+(x_2-2)^2)^{-1} \\ (1+(x_3+1)^2)^{-1} \end{bmatrix}$$

and use the generalized SSD (5.3) to compute the index-1 and index-2 saddle points of this dynamical system with the initial conditions

$$x_0=(-1,1,0)^\top,\quad v_0=(-1,0,0)^\top,\quad x_0=(-1,1,0)^\top,$$
$$v_{1,0}=\frac{1}{\sqrt{2}}(-1,1,0)^\top,\quad v_{2,0}=\frac{1}{\sqrt{2}}(1,1,0)^\top,$$

respectively. Numerical results are presented in Tables 3-4, which again show the first-order accuracy of the scheme (5.3).

6.3 Second-order scheme for gradient and non-gradient systems

We test the convergence rates of the Richardson extrapolation technique proposed in

Section 4.2 by the same examples in Sections 6.1-6.2 and numerical results are presented in Tables 5-8, which show the second-order accuracy of the Richardson extrapolation.

Table 3: Convergence of (5.3) for finding an index-1 saddle point.

$1/\tau$	$\max_n \|e_n^x\|$	CR	$\max_n \|e_n^v\|$	CR
2^5	4.95E-02		9.32E-03	
2^6	2.50E-02	0.98	4.64E-03	1.00
2^7	1.25E-02	1.00	2.30E-03	1.01
2^8	6.19E-03	1.02	1.13E-03	1.02

Table 4: Convergence of (5.3) for finding an index-2 saddle point.

$1/\tau$	$\max_n \|e_n^x\|$	CR	$\max_n \|e_n^v\|$	CR
2^5	3.00E-02		1.02E-02	
2^6	1.50E-02	1.01	5.08E-03	1.00
2^7	7.42E-03	1.01	2.52E-03	1.01
2^8	3.65E-03	1.02	1.24E-03	1.02

Table 5: Convergence of Richardson extrapolation for finding an index-1 saddle point of the gradient system.

$1/\tau$	$\max_n \|e_n^{R,x}\|$	CR	$\max_n \|e_n^{R,v}\|$	CR
2^5	1.45E-03		5.49E-04	
2^6	3.46E-04	2.07	1.34E-04	2.03
2^7	8.43E-05	2.04	3.31E-05	2.02
2^8	2.08E-05	2.02	8.22E-06	2.01

Table 6: Convergence of Richardson extrapolation for finding an index-2 saddle point of the gradient system.

$1/\tau$	$\max_n \|e_n^{R,x}\|$	CR	$\max_n \|e_n^{R,v}\|$	CR
2^5	3.39E-04		9.79E-04	
2^6	8.32E-05	2.03	2.41E-04	2.02
2^7	2.06E-05	2.01	5.97E-05	2.01
2^8	5.13E-06	2.01	1.49E-05	2.01

Table 7: Convergence of Richardson extrapolation for finding an index-1 saddle point of the non-gradient system.

$1/\tau$	$\max_n \|e_n^{R,x}\|$	CR	$\max_n \|e_n^{R,v}\|$	CR
2^5	9.54E-04		1.43E-04	
2^6	2.45E-04	1.96	3.52E-05	2.02
2^7	6.20E-05	1.98	8.71E-06	2.01
2^8	1.56E-05	1.99	2.17E-06	2.01

Table 8: Convergence of Richardson extrapolation for finding an index-2 saddle point of the non-gradient system.

$1/\tau$	$\max_n \|e_n^{R,x}\|$	CR	$\max_n \|e_n^{R,v}\|$	CR
2^5	1.43E-04		1.53E-04	
2^6	3.55E-05	2.01	3.86E-05	1.99
2^7	8.87E-06	2.00	9.69E-06	1.99
2^8	2.21E-06	2.00	2.42E-06	2.00

7 Conclusions

Finding the saddle points of complicated systems has attracted an increasing interest in recent decades. In particular, the SSD serves as an efficient numerical algorithm to compute any-index saddle points and has been widely used to construct the solution landscapes of varied energy and dynamical systems. In this paper we prove the boundedness of the exact solutions and optimal-order error estimates of the numerical discretization to the SSD with respect to the time step size. We overcome the main difficulties of dealing with the local Lipschitz assumptions, the dimer approximation, and the strong nonlinearity of the SSD. We further employ the Richardson extrapolation to obtain the approximate solution with second-order accuracy. The derived analysis and numerical results provide mathematical and numerical supports for the computations of saddle points. In future works, we will investigate how to relax or eliminate the restrictions like (2.8) on the parameters to improve the analysis.

Acknowledgments

This work was partially supported by the National Natural Science Foundation of China (Grant Nos. 12288101, 12225102, 12050002), by the National Key R&D Program of China (Grant No. 2021YFF1200500), by the International Postdoctoral Exchange Fellowship Program (Talent-Introduction Program) (Grant No. YJ20210019), and by the China Postdoctoral Science Foundation (Grant Nos. 2021TQ0017, 2021M700244).

References

[1] E. L. Allgower and K. Georg, *Introduction to Numerical Continuation Methods*, SIAM, 2003.
[2] C. Chen and Z. Xie, *Search extension method for multiple solutions of a nonlinear problem*, Comput. Math. Appl., 47:327–343, 2004.
[3] V. Comincioli, *Analisi Numerica Metodi Modelli Applicazioni*, McGraw-Hill Libri Italia, 1995.
[4] J. Doye and D. Wales, *Saddle points and dynamics of Lennard-Jones clusters, solids, and supercooled liquids*, J. Chem. Phys., 116:3777–3788, 2002.
[5] W. E, E. Vanden-Eijnden, *Transition-path theory and path-finding algorithms for the study of rare events*, Annu. Rev. Phys. Chem., 61:391–420, 2010.
[6] W. E and X. Zhou, *The gentlest ascent dynamics*, Nonlinearity, 24:1831–1842, 2011.

[7] P. E. Farrell, Á. Birkisson, and S. W. Funke, *Deflation techniques for finding distinct solutions of nonlinear partial differential equations*, SIAM J. Sci. Comput., 37:A2026–A2045, 2015.
[8] W. Gao, J. Leng, and X. Zhou, *An iterative minimization formulation for saddle point search*, SIAM J. Numer. Anal., 53:1786–1805, 2015.
[9] N. Gould, C. Ortner, and D. Packwood, *A dimer-type saddle search algorithm with preconditioning and linesearch*, Math. Comp., 85:2939–2966, 2016.
[10] W. Grantham, *Gradient transformation trajectory following algorithms for determining stationary min-max saddle points*, in: Advances in Dynamic Game Theory, Ann. Internat. Soc. Dynam. Games, 9: 639–657, 2007.
[11] Y. Han, Y. Hu, P. Zhang, and L. Zhang, *Transition pathways between defect patterns in confined nematic liquid crystals*, J. Comput. Phys., 396:1–11, 2019.
[12] Y. Han, Z. Xu, A. Shi, and L. Zhang, *Pathways connecting two opposed bilayers with a fusion pore: a molecularly-informed phase field approach*, Soft Matter., 16:366–374, 2020.
[13] Y. Han, J. Yin, Y. Hu, A. Majumdar, and L. Zhang, *Solution landscapes of the simplified Ericksen-Leslie model and its comparison with the reduced Landau-de Gennes model*, Proc. Math. Phys. Eng. Sci., 477:2021.0458, 2021.
[14] Y. Han, J. Yin, P. Zhang, A. Majumdar, and L. Zhang, *Solution landscape of a reduced Landau–de Gennes model on a hexagon*, Nonlinearity, 34:2048–2069, 2021.
[15] D. Heidrich and W. Quapp, *Saddle points of index* 2 *on potential energy surfaces and their role in theoretical reactivity investigations*, Theor. Chim. Acta, 70: 89–98, 1986.
[16] G. Henkelman, H. Jónsson, *A dimer method for finding saddle points on high dimensional potential surfaces using only first derivatives*, J. Chem. Phys., 111:7010–7022, 1999.
[17] A. Levitt and C. Ortner, *Convergence and cycling in Walker-type saddle search algorithms*, SIAM J. Numer. Anal., 55:2204–2227, 2017.
[18] Y. Li and J. Zhou, *A minimax method for finding multiple critical points and its applications to semilinear PDEs*, SIAM J. Sci. Comput., 23:840–865, 2001.
[19] D. Mehta, *Finding all the stationary points of a potential-energy landscape via numerical polynomial-homotopy-continuation method*, Phys. Rev. E, 84:025702, 2011.
[20] J. W. Milnor, *Morse Theory*, Princeton University Press, 1963.
[21] Q. Nie, L. Qiao, Y. Qiu, L. Zhang, and W. Zhao, *Noise control and utility: from regulatory network to spatial patterning*, Sci. China Math., 63:425–440, 2020.
[22] A. Quarteroni, R. Sacco, and F. Saleri, *Numerical Mathematics*. in: Texts in Applied Mathematics 37, 2007.
[23] W. Wang, L. Zhang, P. Zhang, *Modelling and computation of liquid crystals*, Acta Numer., 30:765–851, 2021.
[24] Z. Xu, Y. Han, J. Yin, B. Yu, Y. Nishiura, L. Zhang, *Solution landscapes of the diblock copolymer-homopolymer model under two-dimensional confinement*, Phys. Rev. E, 104:014505, 2021.
[25] J. Yin, K. Jiang, A.-C. Shi, P. Zhang, and L. Zhang, *Transition pathways connecting crystals and quasicrystals*, Proc. Natl. Acad. Sci., 118:e2106230118, 2021.
[26] J. Yin, Y. Wang, J. Chen, P. Zhang, and L. Zhang, *Construction of a pathway map on a complicated energy landscape*, Phys. Rev. Lett., 124:090601, 2020.
[27] J. Yin, B. Yu, and L. Zhang, *Searching the solution landscape by generalized high-index saddle dynamics*, Sci. China Math., 64:1801, 2021.
[28] 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:A3576–A3595, 2019.
[29] J. Yin, L. Zhang, and P. Zhang, *Solution landscape of the Onsager model identifies non-axisymmetric critical points*, Physica D: Nonlinear Phenomena, 430:133081, 2022.

[30] B. Yu, X. Zheng, P. Zhang, and L. Zhang, *Computing solution landscape of nonlinear space-fractional problems via fast approximation algorithm*, J. Comput. Phys., 468:111513, 2022.
[31] J. Zhang and Q. Du, *Shrinking dimer dynamics and its applications to saddle point search*, SIAM J. Numer. Anal., 50:1899–1921, 2012.
[32] L. Zhang, L. Chen, and Q. Du, *Morphology of critical nuclei in solid-state phase transformations*, Phys. Rev. Lett., 98:265703, 2007.
[33] L. Zhang, L. Chen, and Q. Du, *Simultaneous prediction of morphologies of a critical nucleus and an equilibrium precipitate in solids*, Commun. Comput. Phys., 7:674–682, 2010.
[34] L. Zhang, Q. Du, and Z. Zheng, *Optimization-based shrinking dimer method for finding transition states*, SIAM J. Sci. Comput., 38:A528–A544, 2016.
[35] L. Zhang, W. Ren, A. Samanta, and Q. Du, *Recent developments in computational modelling of nucleation in phase transformations*, npj Comput. Mater., 2:16003, 2016.
[36] L. Zhang, P. Zhang, and X. Zheng, *Error estimates of Euler discretization to high-index saddle dynamics*, SIAM J. Numer. Anal., 60:2925–2944, 2022.