Chin. Ann. Math. Ser. B
44(5), 2023, 765–780
DOI: 10.1007/s11401-023-0043-8

Chinese Annals of Mathematics, Series B

Error Estimate for Semi-implicit Method of Sphere-Constrained High-Index Saddle Dynamics*

Lei ZHANG[1] Pingwen ZHANG[2] Xiangcheng ZHENG[3]

Abstract The authors prove error estimates for the semi-implicit numerical scheme of sphere-constrained high-index saddle dynamics, which serves as a powerful instrument in finding saddle points and constructing the solution landscapes of constrained systems on the high-dimensional sphere. Due to the semi-implicit treatment and the novel computational procedure, the orthonormality of numerical solutions at each time step could not be fully employed to simplify the derivations, and the computations of the state variable and directional vectors are coupled with the retraction, the vector transport and the orthonormalization procedure, which significantly complicates the analysis. They address these issues to prove error estimates for the proposed semi-implicit scheme and then carry out numerical experiments to substantiate the theoretical findings.

Keywords Saddle point, Constrained saddle dynamics, Solution landscape, Semi-implicit, Numerical analysis
2000 MR Subject Classification 37M05, 37N30, 65L20

1 Introduction

High-index saddle dynamics (see [25]) attracts increasing attentions in the last few years due to its capability of effectively finding multiple high-index saddle points of complex systems (see [5, 29–30]). Here the index of saddle point refers to the Morse index characterized by the maximal dimension of a subspace on which its Hessian operator is negative definite (see [17]). In particular, the high-index saddle dynamics could be further combined with the downward and upward algorithms (see [24]) to construct the solution landscape, the pathway map consisting of all stationary points and their connections (see [19]), that arises several successful applications (see [10–11, 22–23, 26–28]). In practical problems such as the Thomson problem (see [18]) and the Bose-Einstein Condensation (see [2]), the state variable is constrained on a high-dimensional sphere, which leads to the more complicated sphere-constrained high-index saddle dynamics for treating the sphere-constrained problems.

Manuscript received April 24, 2023. Revised May 18, 2023.
[1]Beijing International Center for Mathematical Research, Center for Machine Learning Research, Center for Quantitative Biology, Peking University, Beijing 100871, China. E-mail: zhangl@math.pku.edu.cn
[2]School of Mathematics and Statistics, Wuhan University, Wuhan 430072, China; School of Mathematical Sciences, Laboratory of Mathematics and Applied Mathematics, Peking University, Beijing 100871, China. E-mail: pzhang@pku.edu.cn
[3]School of Mathematics, Shandong University, Jinan 250100, China. E-mail: xzheng@sdu.edu.cn
*This work was supported by the National Natural Science Foundation of China (Nos. 12225102, 12050002, 12288101, 12301555), the National Key R&D Program of China (No. 2021YFF1200500) and the Taishan Scholars Program of Shandong Province.

There exist extensive works about numerical analysis to algorithms of finding index-1 saddle points (see [1, 3–4, 6–9, 13–14, 16, 20]), while the corresponding analysis for high-index saddle point searchers is rare. In [31], an explicit scheme for the unconstrained high-index saddle dynamics was rigorously analyzed by overcoming the difficulties caused by the coupling of solutions and the (nonlinear) orthonormal procedure of directional vectors in the numerical scheme. The developed method was then extended to prove error estimates for the explicit scheme of the sphere-constrained high-index saddle dynamics by accounting for the more complex dynamical form and additional operations in the numerical scheme such as the retraction and vector transport in order to maintain the manifold constraint (see [21]). To improve the numerical stability, a semi-implicit numerical scheme for the unconstrained high-index saddle dynamics was recently analyzed in [15], and various numerical experiments demonstrated that comparing with the explicit scheme, the semi-implicit method could improve the convergence behavior, admit much larger step size and reduce the number of queries for the model.

The current work is a continuation of the aforementioned sequence of investigations for numerical analysis of high-index saddle dynamics, which will develop and analyze the semi-implicit numerical method for the sphere-constrained high-index saddle dynamics. To achieve this goal, not only do we need to accommodate the complicated nonlinear forms of this dynamical system, the retraction of the state variable, the vector transport and orthonormalization of the directional vectors due to the manifold constraint, but novel techniques are required to overcome the difficulties caused by the semi-implicit treatment. The derived results provide theoretical supports for the numerical accuracy of discretization of sphere-constrained high-index saddle dynamics and construction of solution landscapes for complex systems.

The rest of the paper is organized as follows: In Section 2 we present formulations of the sphere-constrained high-index saddle dynamics and its semi-implicit numerical scheme. In Section 3 we prove several auxiliary estimates, based on which we derive error estimates for the semi-implicit scheme of sphere-constrained high-index saddle dynamics in Section 4. Numerical experiments are performed in Section 5 to substantiate the theoretical findings, and we address concluding remarks in the last section.

2 Problem Formulation and Semi-Implicit Scheme

In this section we propose the semi-implicit numerical scheme of the sphere-constrained high-index saddle dynamics. Let $E(x)$ be the energy function with $x \in \mathbb{R}^d$, and define $F(x) = -\nabla E(x)$ and $J(x) = -\nabla^2 E(x)$ with $J(x) = J(x)^{\mathrm{T}}$. The high-index saddle dynamics for an index-k saddle point of $E(x)$ constrained on the unit sphere S^{d-1} was developed in [21]:

$$\begin{cases} \dfrac{\mathrm{d}x}{\mathrm{d}t} = \Big(I - xx^{\mathrm{T}} - 2\displaystyle\sum_{j=1}^{k} v_j v_j^{\mathrm{T}}\Big) F(x); \\ \dfrac{\mathrm{d}v_i}{\mathrm{d}t} = \Big(I - xx^{\mathrm{T}} - v_i v_i^{\mathrm{T}} - 2\displaystyle\sum_{j=1}^{i-1} v_j v_j^{\mathrm{T}}\Big) J(x) v_i + x v_i^{\mathrm{T}} F(x) \end{cases} \tag{2.1}$$

for $1 \le i \le k$, equipped with the initial conditions

$$x(0) = x_0 \in S^{d-1}, \quad v_i(0) = v_{i,0}$$
$$\text{such that} \quad v_{i,0}^{\mathrm{T}} v_{j,0} = \delta_{ij} \quad \text{and} \quad x_0^{\mathrm{T}} v_{i,0} = 0 \quad \text{for } 1 \le i, j \le k.$$

Here x represents a position variable and $\{v_i\}_{i=1}^k$ are k directional variables. It was proved in [21] that a linearly stable steady state of (2.1) is an index-k saddle point, and the solutions x and $\{v_i\}_{i=1}^k$ to the dynamics (2.1) satisfy for $t>0$,

$$x(t)\in S^{d-1},\quad v_i(t)^{\mathrm{T}}x(t)=0,\quad v_i(t)^{\mathrm{T}}v_j(t)=\delta_{ij},\quad 1\le i,j\le k. \tag{2.2}$$

Throughout the paper we apply the following assumptions.

Assumption A The $F(x)$ could be represented as a sum of the linear part $\mathcal{L}x$ and the nonlinear part $\mathcal{N}(x)$, that is, $F(x)=\mathcal{L}x+\mathcal{N}(x)$, and there exists a constant $L>0$ such that the following linearly growth and Lipschitz conditions hold under the standard l^2 norm $\|\cdot\|$ of a vector or a matrix

$$\max\{\|J(x_2)-J(x_1)\|,\|\mathcal{L}x_2-\mathcal{L}x_1\|,\|\mathcal{N}(x_2)-\mathcal{N}(x_1)\|\}\le L\|x_2-x_1\|,$$
$$\max\{\|\mathcal{L}x\|,\|\mathcal{N}(x)\|\}\le L(1+\|x\|),\quad x,x_1,x_2\in\mathbb{R}^d.$$

To derive the semi-implicit discretization, let $0=t_0<t_1<\cdots<t_N=T$ be the uniform partition of $[0,T]$ with the step size $\tau=T/N$, and let $\{x_n,v_{i,n}\}_{n=0}^N$ be the numerical solution of (2.1). Then we discretize the first-order derivative by the Euler scheme and treat the linear and nonlinear parts on the right-hand side of (2.1) via the implicit and explicit manner, respectively, to obtain the semi-implicit scheme of (2.1) for $1\le n\le N$ as follows:

$$\begin{cases}\widetilde{x}_n=x_{n-1}+\tau\Big(I-2\displaystyle\sum_{j=1}^{k}v_{j,n-1}v_{j,n-1}^{\mathrm{T}}\Big)(\mathcal{L}\widetilde{x}_n+\mathcal{N}(x_{n-1}))\\ \qquad -\tau x_{n-1}x_{n-1}^{\mathrm{T}}(\mathcal{L}x_{n-1}+\mathcal{N}(x_{n-1})),\\ x_n=\dfrac{\widetilde{x}_n}{\|\widetilde{x}_n\|};\\ \left.\begin{array}{l}\widetilde{v}_{i,n}=v_{i,n-1}+\tau\Big(I-x_nx_n^{\mathrm{T}}-2\displaystyle\sum_{j=1}^{i-1}v_{j,n}v_{j,n}^{\mathrm{T}}\Big)J(x_n)\widetilde{v}_{i,n}\\ \qquad -\tau v_{i,n-1}v_{i,n-1}^{\mathrm{T}}J(x_n)v_{i,n-1}+\tau x_n\widetilde{v}_{i,n}^{\mathrm{T}}F(x_n),\\ \widehat{v}_{i,n}=\widetilde{v}_{i,n}-\widetilde{v}_{i,n}^{\mathrm{T}}x_nx_n,\\ v_{i,n}=\mathrm{GS}(\widehat{v}_{i,n},\{v_{j,n}\}_{j=1}^{i-1}),\end{array}\right\}\ 1\le i\le k.\end{cases} \tag{2.3}$$

Here the Gram-Schmidt orthonormalization function $\mathrm{GS}(\widehat{v}_{i,n},\{v_{j,n}\}_{j=1}^{i-1})$ generates the normalized vector $v_{i,n}$ from $\widehat{v}_{i,n}$ that is orthogonal with $\{v_{j,n}\}_{j=1}^{i-1}$, that is,

$$v_{i,n}=\mathcal{N}\Big(\widehat{v}_{i,n}-\sum_{j=1}^{i-1}(\widehat{v}_{i,n}^{\mathrm{T}}v_{j,n})v_{j,n}\Big):=\frac{1}{Y_{i,n}}\Big(\widehat{v}_{i,n}-\sum_{j=1}^{i-1}(\widehat{v}_{i,n}^{\mathrm{T}}v_{j,n})v_{j,n}\Big),$$

where $\mathcal{N}$ is the normalized operator and the normalized factor $Y_{i,n}$ is thus defined as

$$Y_{i,n}:=\Big\|\widehat{v}_{i,n}-\sum_{j=1}^{i-1}(\widehat{v}_{i,n}^{\mathrm{T}}v_{j,n})v_{j,n}\Big\|=\Big(\|\widehat{v}_{i,n}\|^2-\sum_{j=1}^{i-1}(\widehat{v}_{i,n}^{\mathrm{T}}v_{j,n})^2\Big)^{\frac{1}{2}}.$$

The first and the third schemes in (2.3) are semi-implicit discretizations of the equations of x and v_i in (2.1), respectively. The second equation of (2.3) represents the retraction in order to ensure that $x_n \in S^{d-1}$. The last two schemes, which stand for the vector transport and the Gram-Schmidt orthonormalization procedure, respectively, aim to ensure the rest properties of (2.2), that is,

$$v_{i,n}^{\mathrm{T}} x_n = 0, \quad v_{i,n}^{\mathrm{T}} v_{j,n} = \delta_{ij}, \quad 1 \le i, j \le k, \quad 0 \le n \le N. \tag{2.4}$$

Different from the explicit scheme presented in [33], where all variables on the right-hand side of (2.3) take their values at the previous time step t_{n-1}, the orthonormal property of the vectors $\{v_{i,n-1}\}_{i=1}^k$ at the time step t_{n-1} could no longer be fully employed in (2.3) to facilitate the numerical analysis as performed in [33] due to the semi-implicit treatment, which complicates the error estimate. On the other hand, in the explicit scheme the vectors $\{\widetilde{v}_{i,n}\}_{i=1}^k$ are firstly solved, and then their orthonormalization are independently performed. In the semi-implicit scheme (2.3), the computational strategy is quite different in that the last three schemes of directional vectors in (2.3) are sequentially solved for $1 \le i \le k$. In this way, the newly computed orthonormalized vectors $\{v_{j,n}\}_{j=1}^{i-1}$ at the current time step t_n are involved in the scheme of $\widetilde{v}_{i,n}$, which could be more appropriate than invoking the vectors at the previous time step in the explicit scheme. However, this computational strategy leads to the coupling of the schemes of directional vectors, the vector transport and the orthonormalization procedure, which makes the numerical analysis more challenging.

Concerning these difficulties, we derive novel analysis methods to carry out error estimates in subsequent sections. Throughout the paper we use Q to denote a generic positive constant that may assume different values at different occurrences.

3 Auxiliary Estimates

We prove several properties of the numerical solutions to support the error estimates. By $\|x_n\| = \|v_{i,n}\| = 1$ for $1 \le i \le k$ and $1 \le n \le N$, we could apply the Assumption A to derive from the first and the third equations of the scheme (2.3) that

$$\max\{\|\widetilde{x}_n\|, \|\widetilde{v}_{1,n}\|, \cdots, \|\widetilde{v}_{k,n}\|\} \le Q \tag{3.1}$$

for $1 \le n \le N$ for τ small enough, which will be frequently used in the analysis.

Lemma 3.1 *Under Assumption* A*, the following estimate holds for τ small enough*:

$$\|x_n - \widetilde{x}_n\| \le Q\tau^2, \quad 1 \le n \le N; \tag{3.2}$$

$$\|\widehat{v}_{i,n} - \widetilde{v}_{i,n}\| = |\widetilde{v}_{i,n}^{\mathrm{T}} x_n| \le Q\tau^2, \quad 1 \le i \le k, \quad 1 \le n \le N. \tag{3.3}$$

Proof We employ the first equation of (2.3) to get

$$\begin{aligned} \|\widetilde{x}_n - x_{n-1}\| = \Big\| &\tau\Big(I - 2\sum_{j=1}^k v_{j,n-1} v_{j,n-1}^{\mathrm{T}}\Big)(\mathcal{L}\widetilde{x}_n + \mathcal{N}(x_{n-1})) \\ &- \tau x_{n-1} x_{n-1}^{\mathrm{T}}(\mathcal{L}x_{n-1} + \mathcal{N}(x_{n-1}))\Big\| \le Q\tau. \end{aligned} \tag{3.4}$$

We then apply this to rewrite the first equation of (2.3) as

$$\widetilde{x}_n = x_{n-1} + \tau\Big(I - 2\sum_{j=1}^k v_{j,n-1} v_{j,n-1}^{\mathrm{T}}\Big)(\mathcal{L}\widetilde{x}_n + \mathcal{N}(x_{n-1}))$$

$$
\begin{aligned}
&- \tau x_{n-1}x_{n-1}^{\mathrm{T}}(\mathcal{L}x_{n-1} + \mathcal{N}(x_{n-1}))\\
&= x_{n-1} + \tau\Big(I - x_{n-1}x_{n-1}^{\mathrm{T}} - 2\sum_{j=1}^{k} v_{j,n-1}v_{j,n-1}^{\mathrm{T}}\Big)(\mathcal{L}\widetilde{x}_n + \mathcal{N}(x_{n-1}))\\
&\quad + \tau x_{n-1}x_{n-1}^{\mathrm{T}}\mathcal{L}(\widetilde{x}_n - x_{n-1})\\
&= x_{n-1} + \tau\Big(I - x_{n-1}x_{n-1}^{\mathrm{T}} - 2\sum_{j=1}^{k} v_{j,n-1}v_{j,n-1}^{\mathrm{T}}\Big)(\mathcal{L}\widetilde{x}_n + \mathcal{N}(x_{n-1})) + O(\tau^2). \qquad (3.5)
\end{aligned}
$$

We multiply x_{n-1}^{T} on both sides of this equation and use (2.4) to obtain

$$
x_{n-1}^{\mathrm{T}}\widetilde{x}_n = 1 + O(\tau^2).
$$

We then multiply $\widetilde{x}_n^{\mathrm{T}}$ on both sides of (3.5) and use $x_{n-1}^{\mathrm{T}}v_{j,n-1} = 0$ for $1 \le j \le k$ and $x_{n-1}^{\mathrm{T}}\widetilde{x}_n = 1 + O(\tau^2)$ to obtain

$$
\begin{aligned}
\|\widetilde{x}_n\|^2 &= 1 + \tau\Big(\widetilde{x}_n^{\mathrm{T}} - x_{n-1}^{\mathrm{T}} - 2\sum_{j=1}^{k} \widetilde{x}_n^{\mathrm{T}}v_{j,n-1}v_{j,n-1}^{\mathrm{T}}\Big)(\mathcal{L}\widetilde{x}_n + \mathcal{N}(x_{n-1})) + O(\tau^2)\\
&= 1 + \tau(\widetilde{x}_n - x_{n-1})^{\mathrm{T}}\Big(I - 2\sum_{j=1}^{k} v_{j,n-1}v_{j,n-1}^{\mathrm{T}}\Big)(\mathcal{L}\widetilde{x}_n + \mathcal{N}(x_{n-1})) + O(\tau^2),
\end{aligned}
$$

which, together with Assumption A and the norm-preserving property of the Householder matrix in the above equation, yields

$$
\begin{aligned}
|\|\widetilde{x}_n\|^2 - 1| &\le \tau\|\widetilde{x}_n - x_{n-1}\|\|\mathcal{L}\widetilde{x}_n + \mathcal{N}(x_{n-1})\| + O(\tau^2)\\
&\le Q\tau\|\widetilde{x}_n - x_{n-1}\| + O(\tau^2).
\end{aligned}
$$

Combining this equation and (3.4) we obtain

$$
|\|\widetilde{x}_n\|^2 - 1| \le Q\tau^2,
$$

which in turn leads to $|\|\widetilde{x}_n\| - 1| \le Q\tau^2$. We apply this to reach (3.2):

$$
\|x_n - \widetilde{x}_n\| = \Big\|\frac{\widetilde{x}_n}{\|\widetilde{x}_n\|}(1 - \|\widetilde{x}_n\|)\Big\| = |1 - \|\widetilde{x}_n\|| \le Q\tau^2.
$$

To derive (3.3), we combine (3.2) and (3.4) to obtain

$$
\|x_n - x_{n-1}\| \le \|x_n - \widetilde{x}_n\| + \|\widetilde{x}_n - x_{n-1}\| \le Q\tau. \qquad (3.6)
$$

From the forth equation of (2.3) we apply $\|x_n\| = 1$ to obtain

$$
\|\widehat{v}_{i,n} - \widetilde{v}_{i,n}\| = |\widetilde{v}_{i,n}^{\mathrm{T}}x_n|. \qquad (3.7)
$$

Furthermore, the relation $|\|\widetilde{x}_n\| - 1| \le Q\tau^2$ leads to $\|\widetilde{x}_n\| \ge 1 - Q\tau^2 \ge \frac{1}{2}$ for τ small enough. Then we multiply the scheme of $\widetilde{v}_{i,n}$ in (2.3) and the reformulated scheme of $\widetilde{x}_n$ in (3.5) to get

$$
x_n^{\mathrm{T}}\widetilde{v}_{i,n} = \frac{1}{\|\widetilde{x}_n\|}\Big[x_{n-1} + \tau\Big(I - x_{n-1}x_{n-1}^{\mathrm{T}} - 2\sum_{j=1}^{k} v_{j,n-1}v_{j,n-1}^{\mathrm{T}}\Big)
$$

$$
\begin{aligned}
&\cdot (\mathcal{L}\widetilde{x}_n + \mathcal{N}(x_{n-1})) + O(\tau^2)\Big]^{\mathrm{T}} \\
&\Big[v_{i,n-1} + \tau\Big(I - x_n x_n^{\mathrm{T}} - 2\sum_{j=1}^{i-1} v_{j,n}v_{j,n}^{\mathrm{T}}\Big)J(x_n)\widetilde{v}_{i,n} \\
&- \tau v_{i,n-1}v_{i,n-1}^{\mathrm{T}}J(x_n)v_{i,n-1} + \tau x_n \widetilde{v}_{i,n}^{\mathrm{T}}F(x_n)\Big] \\
=\ &\frac{\tau}{\|\widetilde{x}_n\|}\Big[\Big(x_{n-1}^{\mathrm{T}} - x_{n-1}^{\mathrm{T}}x_n x_n^{\mathrm{T}} - 2\sum_{j=1}^{i-1} x_{n-1}^{\mathrm{T}}v_{j,n}v_{j,n}^{\mathrm{T}}\Big)J(x_n)\widetilde{v}_{i,n} \\
&+ x_{n-1}^{\mathrm{T}}x_n\widetilde{v}_{i,n}^{\mathrm{T}}F(x_n) - v_{i,n-1}^{\mathrm{T}}(\mathcal{L}\widetilde{x}_n + \mathcal{N}(x_{n-1}))\Big] + O(\tau^2), \qquad (3.8)
\end{aligned}
$$

where we briefly write the second-order terms of τ as $O(\tau^2)$. We apply the splittings

$$
x_{n-1}^{\mathrm{T}} - x_{n-1}^{\mathrm{T}}x_n x_n^{\mathrm{T}} = (x_{n-1} - x_n)^{\mathrm{T}}(I - x_n x_n^{\mathrm{T}})
$$

and

$$
\begin{aligned}
&x_{n-1}^{\mathrm{T}}x_n\widetilde{v}_{i,n}^{\mathrm{T}}F(x_n) - v_{i,n-1}^{\mathrm{T}}(\mathcal{L}\widetilde{x}_n + \mathcal{N}(x_{n-1})) \\
=\ &x_{n-1}^{\mathrm{T}}x_n\widetilde{v}_{i,n}^{\mathrm{T}}F(x_n) - v_{i,n-1}^{\mathrm{T}}F(x_{n-1}) + v_{i,n-1}^{\mathrm{T}}\mathcal{L}(x_{n-1} - \widetilde{x}_n) \\
=\ &(x_{n-1} - x_n)^{\mathrm{T}}x_n\widetilde{v}_{i,n}^{\mathrm{T}}F(x_n) + (\widetilde{v}_{i,n} - v_{i,n-1})^{\mathrm{T}}F(x_n) \\
&+ v_{i,n-1}^{\mathrm{T}}(F(x_n) - F(x_{n-1})) + v_{i,n-1}^{\mathrm{T}}\mathcal{L}(x_{n-1} - \widetilde{x}_n),
\end{aligned}
$$

to bound the right-hand side of (3.8) as

$$
\begin{aligned}
|x_n^{\mathrm{T}}\widetilde{v}_{i,n}| \le \frac{Q\tau}{\|\widetilde{x}_n\|}\Big[&\|x_n - x_{n-1}\| + \sum_{j=1}^{i-1}|x_{n-1}^{\mathrm{T}}v_{j,n}| \\
&+ \|\widetilde{v}_{i,n} - v_{i,n-1}\| + \|F(x_n) - F(x_{n-1})\| + \|x_{n-1} - \widetilde{x}_n\|\Big] + O(\tau^2). \qquad (3.9)
\end{aligned}
$$

We then invoke the third scheme of (2.3),

$$
\begin{aligned}
\|\widetilde{v}_{i,n} - v_{i,n-1}\| = \Big\|&\tau\Big(I - x_n x_n^{\mathrm{T}} - 2\sum_{j=1}^{i-1} v_{j,n}v_{j,n}^{\mathrm{T}}\Big)J(x_n)\widetilde{v}_{i,n} \\
&- \tau v_{i,n-1}v_{i,n-1}^{\mathrm{T}}J(x_n)v_{i,n-1} + \tau x_n\widetilde{v}_{i,n}^{\mathrm{T}}F(x_n)\Big\| \le Q\tau, \qquad (3.10)
\end{aligned}
$$

as well as $x_{n-1}^{\mathrm{T}}v_{j,n} = (x_{n-1} - x_n)^{\mathrm{T}}v_{j,n}$, $\|\widetilde{x}_n\| \ge \frac{1}{2}$, (3.4), (3.6) and the Lipschitz condition of F in (3.8) to obtain

$$
\begin{aligned}
|\widetilde{v}_{i,n}^{\mathrm{T}}x_n| \le Q\tau\Big[&\|x_n - x_{n-1}\| + \sum_{j=1}^{i-1}\|x_{n-1} - x_n\|\|v_{j,n}\| \\
&+ \|\widetilde{v}_{i,n} - v_{i,n-1}\| + \|x_{n-1} - \widetilde{x}_n\|\Big] + O(\tau^2) \\
\le\ &Q\tau(\|x_{n-1} - x_n\| + \|\widetilde{v}_{i,n} - v_{i,n-1}\| + \|x_{n-1} - \widetilde{x}_n\|) + O(\tau^2) \le Q\tau^2,
\end{aligned}
$$

which completes the proof.

Lemma 3.2 *For $1 \le m < i \le k$ and $1 \le j \le k$, the following estimates hold for τ small enough*:

$$\|\widetilde{v}_{i,n}^{\mathrm{T}}\widetilde{v}_{m,n}\| \le Q\tau\sum_{l=1}^{m}\|\widehat{v}_{l,n}-v_{l,n}\|+Q\tau^2,$$
$$|\|\widetilde{v}_{j,n}\|^2-1| \le Q\tau\sum_{l=1}^{j-1}\|\widehat{v}_{l,n}-v_{l,n}\|+Q\tau^2.$$

Proof From the definitions of $\widetilde{v}_{i,n}$ and $\widetilde{v}_{m,n}$ we have

$$\begin{aligned}\widetilde{v}_{i,n}^{\mathrm{T}}\widetilde{v}_{m,n} &= \tau\Big(v_{m,n-1}^{\mathrm{T}}J(x_n)\widetilde{v}_{i,n}-x_n^{\mathrm{T}}v_{m,n-1}\widetilde{v}_{i,n}J(x_n)^{\mathrm{T}}x_n\\ &\quad-2\sum_{j=1}^{i-1}v_{m,n-1}^{\mathrm{T}}v_{j,n}v_{j,n}^{\mathrm{T}}J(x_n)\widetilde{v}_{i,n}+x_n^{\mathrm{T}}v_{m,n-1}\widetilde{v}_{i,n}^{\mathrm{T}}F(x_n)\\ &\quad+\widetilde{v}_{m,n}^{\mathrm{T}}J(x_n)^{\mathrm{T}}v_{i,n-1}-x_n^{\mathrm{T}}v_{i,n-1}x_n^{\mathrm{T}}J(x_n)\widetilde{v}_{m,n}\\ &\quad-2\sum_{j=1}^{m-1}v_{j,n}^{\mathrm{T}}v_{i,n-1}v_{j,n}^{\mathrm{T}}J(x_n)\widetilde{v}_{m,n}+v_{i,n-1}^{\mathrm{T}}x_n\widetilde{v}_{m,n}^{\mathrm{T}}F(x_n)\Big)+O(\tau^2)\\ &=: \sum_{l=1}^{8}K_l+O(\tau^2).\end{aligned}$$

We apply $x_n^{\mathrm{T}}v_{i,n}=0$ for $1\le i\le k$ and $1\le n\le N$ and (3.6) to bound $K_2+K_4+K_6+K_8$ as

$$\begin{aligned}&\|K_2+K_4+K_6+K_8\|\\ &=\tau\|-x_n^{\mathrm{T}}v_{m,n-1}\widetilde{v}_{i,n}J(x_n)^{\mathrm{T}}x_n+x_n^{\mathrm{T}}v_{m,n-1}\widetilde{v}_{i,n}^{\mathrm{T}}F(x_n)\\ &\quad-x_n^{\mathrm{T}}v_{i,n-1}x_n^{\mathrm{T}}J(x_n)\widetilde{v}_{m,n}+v_{i,n-1}^{\mathrm{T}}x_n\widetilde{v}_{m,n}^{\mathrm{T}}F(x_n)\|\\ &=\tau\|-(x_n-x_{n-1})^{\mathrm{T}}v_{m,n-1}\widetilde{v}_{i,n}J(x_n)^{\mathrm{T}}x_n\\ &\quad+(x_n-x_{n-1})^{\mathrm{T}}v_{m,n-1}\widetilde{v}_{i,n}^{\mathrm{T}}F(x_n)\\ &\quad-(x_n-x_{n-1})^{\mathrm{T}}v_{i,n-1}x_n^{\mathrm{T}}J(x_n)\widetilde{v}_{m,n}\\ &\quad+v_{i,n-1}^{\mathrm{T}}(x_n-x_{n-1})\widetilde{v}_{m,n}^{\mathrm{T}}F(x_n)\|\le Q\tau^2.\end{aligned}$$

We then introduce the following triple splitting:

$$v_{i,n-1}-v_{i,n}=(v_{i,n-1}-\widetilde{v}_{i,n})+(\widetilde{v}_{i,n}-\widehat{v}_{i,n})+(\widehat{v}_{i,n}-v_{i,n}).$$

The first right-hand side term is estimated by (3.10) and the second right-hand side term is bounded by Lemma 3.1, which lead to

$$\|v_{i,n-1}-v_{i,n}\|\le Q\tau+\|\widehat{v}_{i,n}-v_{i,n}\|. \tag{3.11}$$

We invoke this to bound K_7 as

$$\begin{aligned}|K_7| &= \Big|2\tau\gamma\sum_{j=1}^{m-1}v_{j,n}^{\mathrm{T}}v_{i,n-1}v_{j,n}^{\mathrm{T}}J(x_n)\widetilde{v}_{m,n}\Big|\\ &=\Big|2\tau\gamma\sum_{j=1}^{m-1}(v_{j,n}^{\mathrm{T}}-v_{j,n-1}^{\mathrm{T}})v_{i,n-1}v_{j,n}^{\mathrm{T}}J(x_n)\widetilde{v}_{m,n}\Big|\end{aligned}$$

$$\leq Q\tau^2 + Q\tau \sum_{j=1}^{m-1} \|v_{j,n} - \widehat{v}_{j,n}\|.$$

By $v_{m,n}^{\mathrm{T}} v_{j,n} = \delta_{m,j}$ we rewrite K_3 as

$$\begin{aligned}
K_3 &= -2\tau \sum_{j=1}^{i-1} v_{m,n-1}^{\mathrm{T}} v_{j,n} v_{j,n}^{\mathrm{T}} J(x_n) \widetilde{v}_{i,n} \\
&= -2\tau \sum_{j=1}^{i-1} (v_{m,n-1}^{\mathrm{T}} - v_{m,n}^{\mathrm{T}}) v_{j,n} v_{j,n}^{\mathrm{T}} J(x_n) \widetilde{v}_{i,n} - 2\tau v_{m,n}^{\mathrm{T}} J(x_n) \widetilde{v}_{i,n},
\end{aligned} \tag{3.12}$$

which leads to

$$\begin{aligned}
K_1 + K_3 + K_5 = {} & \tau (v_{m,n-1}^{\mathrm{T}} J(x_n) \widetilde{v}_{i,n} - v_{m,n}^{\mathrm{T}} J(x_n) \widetilde{v}_{i,n}) \\
& + \tau (\widetilde{v}_{m,n}^{\mathrm{T}} J(x_n)^{\mathrm{T}} v_{i,n-1} - v_{m,n}^{\mathrm{T}} J(x_n) \widetilde{v}_{i,n}) \\
& - 2\tau \sum_{j=1}^{i-1} (v_{m,n-1}^{\mathrm{T}} - v_{m,n}^{\mathrm{T}}) v_{j,n} v_{j,n}^{\mathrm{T}} J(x_n) \widetilde{v}_{i,n} =: B_1 + B_2 + B_3.
\end{aligned} \tag{3.13}$$

We then use (3.11) to bound B_1 as

$$|B_1| = \tau |(v_{m,n-1}^{\mathrm{T}} - v_{m,n}^{\mathrm{T}}) J(x_n) \widetilde{v}_{i,n}| \leq Q\tau^2 + Q\tau \|\widehat{v}_{m,n} - v_{m,n}\|.$$

B_3 could be estimated similarly:

$$|B_3| = 2\tau \Big| \sum_{j=1}^{i-1} (v_{m,n-1}^{\mathrm{T}} - v_{m,n}^{\mathrm{T}}) v_{j,n} v_{j,n}^{\mathrm{T}} J(x_n) \widetilde{v}_{i,n} \Big| \leq Q\tau^2 + Q\tau \|\widehat{v}_{m,n} - v_{m,n}\|.$$

We then apply the symmetry of $J(x_n)$ and Lemma 3.1 and (3.10) to bound B_2 as

$$\begin{aligned}
|B_2| &= \tau |(\widetilde{v}_{m,n}^{\mathrm{T}} - v_{m,n}^{\mathrm{T}}) J(x_n) v_{i,n-1} + v_{m,n}^{\mathrm{T}} J(x_n)(v_{i,n-1} - \widetilde{v}_{i,n})| \\
&= \tau |(\widetilde{v}_{m,n}^{\mathrm{T}} - \widehat{v}_{m,n}^{\mathrm{T}} + \widehat{v}_{m,n}^{\mathrm{T}} - v_{m,n}^{\mathrm{T}}) J(x_n) v_{i,n-1} \\
&\quad + v_{m,n}^{\mathrm{T}} J(x_n)(v_{i,n-1} - \widetilde{v}_{i,n})| \leq Q\tau^2 + Q\tau \|\widehat{v}_{m,n} - v_{m,n}\|.
\end{aligned}$$

We incorporate the preceding estimates to complete the proof of the first statement of this lemma.

To derive the second statement, we apply the definition of $\widetilde{v}_{j,n}$ in (2.3) to get

$$\begin{aligned}
\|\widetilde{v}_{j,n}\|^2 = {} & 1 + 2\tau \Big(v_{j,n-1}^{\mathrm{T}} - v_{j,n-1}^{\mathrm{T}} x_n x_n^{\mathrm{T}} - 2 \sum_{l=1}^{j-1} v_{j,n-1}^{\mathrm{T}} v_{l,n} v_{l,n}^{\mathrm{T}} \Big) J(x_n) \widetilde{v}_{j,n} \\
& - 2\tau v_{j,n-1}^{\mathrm{T}} J(x_n) v_{j,n-1} + 2\tau v_{j,n-1}^{\mathrm{T}} x_n \widetilde{v}_{j,n}^{\mathrm{T}} F(x_n) + O(\tau^2),
\end{aligned}$$

that is,

$$\begin{aligned}
|\|\widetilde{v}_{j,n}\|^2 - 1| = \Big| & 2\tau v_{j,n-1}^{\mathrm{T}} J(x_n)(\widetilde{v}_{j,n} - v_{j,n-1}) \\
& - \tau \Big(v_{j,n-1}^{\mathrm{T}} (x_n - x_{n-1}) x_n^{\mathrm{T}}
\end{aligned}$$

$$
\begin{aligned}
&+ 2\sum_{l=1}^{j-1} v_{j,n-1}^{\mathrm{T}}(v_{l,n}-v_{l,n-1})v_{l,n}^{\mathrm{T}}\Big) J(x_n)\widetilde{v}_{j,n} \\
&+ 2\tau v_{j,n-1}^{\mathrm{T}}(x_n-x_{n-1})\widetilde{v}_{j,n}^{\mathrm{T}} F(x_n) + O(\tau^2)\Big|.
\end{aligned}
$$

Thus we incorporate (3.6) and (3.10)–(3.11) to get

$$
|\|\widetilde{v}_{j,n}\|^2 - 1| \le Q\tau \sum_{l=1}^{j-1} \|\widehat{v}_{l,n} - v_{l,n}\| + Q\tau^2,
$$

which completes the proof.

Lemma 3.3 *For $1 \le m < i \le k$ and $1 \le j \le k$, the following estimates hold for τ small enough*:

$$
\begin{aligned}
\|\widehat{v}_{i,n}^{\mathrm{T}}\widehat{v}_{m,n}\| &\le Q_0\tau \sum_{l=1}^{m} \|\widehat{v}_{l,n} - v_{l,n}\| + Q_1\tau^2, \\
|\|\widehat{v}_{j,n}\|^2 - 1| &\le Q_2\tau \sum_{l=1}^{j-1} \|\widehat{v}_{l,n} - v_{l,n}\| + Q_3\tau^2.
\end{aligned}
$$

Proof For $1 \le m < i \le k$ we get

$$
\widehat{v}_{m,n}^{\mathrm{T}}\widehat{v}_{i,n} = \widetilde{v}_{m,n}^{\mathrm{T}}\widetilde{v}_{i,n} - x_n^{\mathrm{T}}\widetilde{v}_{i,n}x_n^{\mathrm{T}}\widetilde{v}_{m,n},
$$

which, together with Lemmas 3.1–3.2, leads to

$$
\begin{aligned}
|\widehat{v}_{m,n}^{\mathrm{T}}\widehat{v}_{i,n}| &\le Q\tau \sum_{l=1}^{m} \|\widehat{v}_{l,n} - v_{l,n}\| + Q\tau^2 + Q\tau^4 \\
&\le Q\tau \sum_{l=1}^{m} \|\widehat{v}_{l,n} - v_{l,n}\| + Q\tau^2.
\end{aligned}
$$

We then apply Lemmas 3.1–3.2 to the relation

$$
\begin{aligned}
\|\widehat{v}_{j,n}\|^2 - 1 &= \|\widetilde{v}_{j,n}\|^2 - 2(x_n^{\mathrm{T}}\widetilde{v}_{j,n})^2 + (x_n^{\mathrm{T}}\widetilde{v}_{j,n})^2 - 1 \\
&= \|\widetilde{v}_{j,n}\|^2 - 1 - (x_n^{\mathrm{T}}\widetilde{v}_{j,n})^2
\end{aligned}
$$

to find

$$
\begin{aligned}
|\|\widehat{v}_{j,n}\|^2 - 1| &\le |\|\widetilde{v}_{j,n}\|^2 - 1| + |(x_n^{\mathrm{T}}\widetilde{v}_{j,n})^2| \\
&\le Q\tau \sum_{l=1}^{j-1} \|\widehat{v}_{l,n} - v_{l,n}\| + Q\tau^2 + Q\tau^4 \\
&\le Q\tau \sum_{l=1}^{j-1} \|\widehat{v}_{l,n} - v_{l,n}\| + Q\tau^2,
\end{aligned}
$$

which completes the proof.

4 Numerical Analysis for Semi-Implicit Scheme

We prove error estimate for the semi-implicit scheme (2.3) by performing a multi-variable circulating induction procedure to gradually decouple the quantities of interest.

4.1 Quantification of $\widetilde{v}_{i,n} - v_{i,n}$

For $\overline{G} > Q_3Q_4 + kQ_1$ where Q_1 and Q_3 are introduced in Lemma 3.3 and $Q_4 > 1$ represents the bound of $\{\widetilde{v}_{j,n}\}_{j=1,n=0}^{k,N}$ (see (3.1)), there exists an intermediate constant $G > 0$ such that

$$\overline{G} > Q_3Q_4 + kG \quad \text{and} \quad G > Q_1.$$

In particular, as $Q_4 > 1$, we have $\overline{G} > Q_3$. Then for τ small enough the following inequalities hold:

$$\frac{Q_0\tau k\overline{G} + Q_1 + kG^2\tau^2}{(1 - Q_2\tau^3 k\overline{G} - Q_3\tau^2 - kG^2\tau^4)^{\frac{1}{2}}} \le G, \quad \frac{Q_4(Q_2\tau k\overline{G} + Q_3 + kG^2\tau^2) + kG}{(1 - Q_2\tau^3 k\overline{G} - Q_3\tau^2 - kG^2\tau^4)^{\frac{1}{2}}} \le \overline{G}. \tag{4.1}$$

In subsequent proofs, we always choose sufficiently small step size τ such that the condition (4.1) is satisfied.

Theorem 4.1 *Under the condition* (4.1), *the following estimate holds for* $1 \le n \le N$:

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

Remark 4.1 The $\widetilde{v}_{i,n}$ on the left-hand side of the third equation of (2.3) could be split as

$$\widetilde{v}_{i,n} = v_{i,n} - (v_{i,n} - \widehat{v}_{i,n}) - (\widehat{v}_{i,n} - \widetilde{v}_{i,n}),$$

where the last two right-hand side terms are $O(\tau^2)$ terms according to Lemma 3.1 and this theorem. Thus we reach the following relation that plays a key role in error estimates:

$$\widetilde{v}_{i,n} = v_{i,n} + O(\tau^2). \tag{4.2}$$

Proof We prove this theorem by induction for the following two relations:

$$(\mathbb{A}): \ \max_{m<i\le k} \|\widehat{v}_{i,n}^{\mathrm{T}} v_{m,n}\| \le G\tau^2 \quad \text{for some } 1 \le m \le k-1;$$

$$(\mathbb{B}): \ \|v_{j,n} - \widehat{v}_{j,n}\| \le \overline{G}\tau^2 \quad \text{for some } 1 \le j \le k.$$

We first declare that if

$$(\mathbb{A}) \text{ holds for } 1 \le m \le m^* - 1 \quad \text{and} \quad (\mathbb{B}) \text{ holds for } 1 \le j \le m^* \tag{4.3}$$

for some $1 \le m^* < k - 1$, then

$$(\mathbb{A}) \text{ holds for } m = m^* \quad \text{and} \quad (\mathbb{B}) \text{ holds for } j = m^* + 1. \tag{4.4}$$

To show this, we apply Lemma 3.3 and the induction hypotheses (4.3) to bound $Y_{m^*,n}$ by

$$\begin{aligned} Y_{m^*,n} &= \Big(\|\widehat{v}_{m^*,n}\|^2 - \sum_{j=1}^{m^*-1} (\widehat{v}_{m^*,n}^{\mathrm{T}} v_{j,n})^2\Big)^{\frac{1}{2}} \\ &\in \Big[1 \pm \Big(Q_2\tau \sum_{l=1}^{m^*-1} \|\widehat{v}_{l,n} - v_{l,n}\| + Q_3\tau^2 + (m^*-1)G^2\tau^4\Big)\Big]^{\frac{1}{2}} \\ &\in [1 \pm (Q_2(m^*-1)\overline{G}\tau^3 + Q_3\tau^2 + (m^*-1)G^2\tau^4)]^{\frac{1}{2}}. \end{aligned} \tag{4.5}$$

We then invoke the induction hypotheses (4.3), (4.5), the condition (4.1) and Lemma 3.3 into the expression of $\widehat{v}_{i,n}^{\mathrm{T}} v_{m^*,n}$ to obtain for $m^* < i \leq k$,

$$\begin{aligned}
|\widehat{v}_{i,n}^{\mathrm{T}} v_{m^*,n}| &= \frac{1}{Y_{m^*,n}} \Big| \widehat{v}_{i,n}^{\mathrm{T}} \widehat{v}_{m^*,n} - \sum_{j=1}^{m^*-1} (\widehat{v}_{m^*,n}^{\mathrm{T}} v_{j,n})(\widehat{v}_{i,n}^{\mathrm{T}} v_{j,n}) \Big| \\
&\leq \frac{1}{Y_{m^*,n}} \Big(Q_0 \tau \sum_{l=1}^{m^*} \|\widehat{v}_{l,n} - v_{l,n}\| + Q_1 \tau^2 + (m^*-1) G^2 \tau^4 \Big) \\
&\leq \frac{Q_0 \tau m^* \overline{G} + Q_1 + (m^*-1) G^2 \tau^2}{(1 - Q_2 \tau^3 (m^*-1)\overline{G} - Q_3 \tau^2 - (m^*-1) G^2 \tau^4)^{\frac{1}{2}}} \tau^2 \leq G\tau^2,
\end{aligned}$$

which implies that ($\mathbb{A}$) holds for $m = m^*$. We then use Lemma 3.3 and ($\mathbb{A}$) with $1 \leq m \leq m^*$ to bound $Y_{m^*+1,n}$ in an analogous manner as (4.5):

$$Y_{m^*+1,n} \in [1 \pm (Q_2 m^* \overline{G} \tau^3 + Q_3 \tau^2 + m^* G^2 \tau^4)]^{\frac{1}{2}}, \tag{4.6}$$

which implies

$$|1 - Y_{m^*+1,n}| \leq |1 - Y_{m^*+1,n}^2| \leq Q_2 m^* \overline{G} \tau^3 + Q_3 \tau^2 + m^* G^2 \tau^4.$$

We invoke this and ($\mathbb{A}$) with $1 \leq m \leq m^*$ in $v_{m^*+1,n} - \widehat{v}_{m^*+1,n}$ to get

$$\begin{aligned}
&\|v_{m^*+1,n} - \widehat{v}_{m^*+1,n}\| \\
&= \frac{1}{Y_{m^*+1,n}} \Big\| (1 - Y_{m^*+1,n}) \widehat{v}_{m^*+1,n} - \sum_{j=1}^{m^*} (\widehat{v}_{m^*+1,n}^{\mathrm{T}} v_{j,n}) v_{j,n} \Big\| \\
&\leq \frac{Q_4 (Q_2 \tau m^* \overline{G} + Q_3 + m^* G^2 \tau^2) + m^* G}{(1 - Q_2 \tau^3 m^* \overline{G} - Q_3 \tau^2 - m^* G^2 \tau^4)^{\frac{1}{2}}} \tau^2 \leq \overline{G} \tau^2,
\end{aligned} \tag{4.7}$$

which implies that ($\mathbb{B}$) holds for $j = m^* + 1$. Therefore, the declaration (4.3)–(4.4) is correct and we remain to show that ($\mathbb{A}$) holds for $m = 1$ and ($\mathbb{B}$) holds for $1 \leq j \leq 2$ in order to start the mathematical induction. We apply Lemma 3.3 to obtain

$$\|\widehat{v}_{1,n} - v_{1,n}\| = \Big\| \frac{\widehat{v}_{1,n}}{\|\widehat{v}_{1,n}\|} (\|\widehat{v}_{1,n}\| - 1) \Big\| \leq |\|\widehat{v}_{1,n}\|^2 - 1| \leq Q_3 \tau^2 \leq \overline{G} \tau^2,$$

which is the relation ($\mathbb{B}$) with $j = 1$. Based on this, ($\mathbb{A}$) with $m = 1$ and ($\mathbb{B}$) with $j = 2$ can be proved following exactly the same procedure as (4.5)–(4.7), which completes the proof.

4.2 Error estimate

We prove error estimates for the semi-implicit scheme (2.3) of sphere-constrained high-index saddle dynamics (2.1) by analyzing the following errors:

$$e_n^x := x(t_n) - x_n, \quad e_n^{v_i} := v_i(t_n) - v_{i,n}, \quad 1 \leq n \leq N, \quad 1 \leq i \leq k.$$

Theorem 4.2 *Under Assumption* A, *the following estimate holds for the semi-implicit scheme* (2.3) *for τ sufficiently small*:

$$\max_{1 \leq n \leq N} \{\|e_n^x\|, \|e_n^{v_1}\|, \cdots, \|e_n^{v_k}\|\} \leq Q\tau, \quad 1 \leq n \leq N.$$

Here Q is independent from τ, n and N.

Proof To bound e_n^x, we derive the reference equation from the first equation of (2.1) via the forward Euler discretization

$$\begin{aligned}x(t_n) = x(t_{n-1}) + \tau\Big(I - x(t_{n-1})x(t_{n-1})^{\mathrm{T}} \\ - 2\sum_{j=1}^{k} v_j(t_{n-1})v_j(t_{n-1})^{\mathrm{T}}\Big)F(x(t_{n-1})) + O(\tau^2).\end{aligned}$$

We then apply (3.2) and (3.4) to reformulate (3.5) as

$$\begin{aligned}x_n &= x_{n-1} + (x_n - \widetilde{x}_n) \\ &\quad + \tau\Big(I - x_{n-1}x_{n-1}^{\mathrm{T}} - 2\sum_{j=1}^{k} v_{j,n-1}v_{j,n-1}^{\mathrm{T}}\Big)(\mathcal{L}\widetilde{x}_n + \mathcal{N}(x_{n-1})) + O(\tau^2) \\ &= x_{n-1} + (x_n - \widetilde{x}_n) \\ &\quad + \tau\Big(I - x_{n-1}x_{n-1}^{\mathrm{T}} - 2\sum_{j=1}^{k} v_{j,n-1}v_{j,n-1}^{\mathrm{T}}\Big)F(x_{n-1}) \\ &\quad + \tau\Big(I - x_{n-1}x_{n-1}^{\mathrm{T}} - 2\sum_{j=1}^{k} v_{j,n-1}v_{j,n-1}^{\mathrm{T}}\Big)\mathcal{L}(\widetilde{x}_n - x_{n-1}) + O(\tau^2) \\ &= x_{n-1} + \tau\Big(I - x_{n-1}x_{n-1}^{\mathrm{T}} - 2\sum_{j=1}^{k} v_{j,n-1}v_{j,n-1}^{\mathrm{T}}\Big)F(x_{n-1}) + O(\tau^2).\end{aligned}$$

By this means, the original semi-implicit scheme of x in (2.3) is converted to the explicit scheme to facilitate the analysis. Based on the above two equations, we follow the same derivations in [33, Theorem 4.2] to obtain

$$\|e_n^x\| \le Q\tau \sum_{m=1}^{n-1}\sum_{j=1}^{k} \|e_m^{v_j}\| + Q\tau, \quad 1 \le n \le N. \tag{4.8}$$

To estimate $e_n^{v_i}$, we derive the reference equation from the third equation of (2.1) via the backward Euler discretization for $1 \le i \le k$:

$$\begin{aligned}v_i(t_n) &= v_i(t_{n-1}) + \tau\Big(I - x(t_n)x(t_n)^{\mathrm{T}} - v_i(t_n)v_i(t_n)^{\mathrm{T}} \\ &\quad - 2\sum_{j=1}^{i-1} v_j(t_n)v_j(t_n)^{\mathrm{T}}\Big)J(x(t_n))v_i(t_n) + \tau x(t_n)v_i(t_n)^{\mathrm{T}}F(x(t_n)) + O(\tau^2) \\ &= v_i(t_{n-1}) + \tau\Big(I - x(t_n)x(t_n)^{\mathrm{T}} - 2\sum_{j=1}^{i-1} v_j(t_n)v_j(t_n)^{\mathrm{T}}\Big)J(x(t_n))v_i(t_n) \\ &\quad - \tau v_i(t_{n-1})v_i(t_{n-1})^{\mathrm{T}}J(x(t_n))v_i(t_{n-1}) \\ &\quad + \tau x(t_n)v_i(t_n)^{\mathrm{T}}F(x(t_n)) + O(\tau^2) + \mathcal{A}_n,\end{aligned}$$

where

$$\begin{aligned}\mathcal{A}_n &= \tau(v_i(t_n)v_i(t_n)^{\mathrm{T}}J(x(t_n))v_i(t_n) \\ &\quad - v_i(t_{n-1})v_i(t_{n-1})^{\mathrm{T}}J(x(t_n))v_i(t_{n-1})) = O(\tau^2).\end{aligned}$$

We then apply (4.2) to rewrite the third scheme of (2.3) as

$$\begin{aligned} v_{i,n} &= v_{i,n-1} + (v_{i,n} - \widetilde{v}_{i,n}) \\ &\quad + \tau\Big(I - x_n x_n^{\mathrm{T}} - 2\sum_{j=1}^{i-1} v_{j,n} v_{j,n}^{\mathrm{T}}\Big) J(x_n)(v_{i,n} + O(\tau^2)) \\ &\quad - \tau v_{i,n-1} v_{i,n-1}^{\mathrm{T}} J(x_n) v_{i,n-1} + \tau x_n (v_{i,n} + O(\tau^2))^{\mathrm{T}} F(x_n) \\ &= v_{i,n-1} + \tau\Big(I - x_n x_n^{\mathrm{T}} - 2\sum_{j=1}^{i-1} v_{j,n} v_{j,n}^{\mathrm{T}}\Big) J(x_n) v_{i,n} \\ &\quad - \tau v_{i,n-1} v_{i,n-1}^{\mathrm{T}} J(x_n) v_{i,n-1} + \tau x_n v_{i,n}^{\mathrm{T}} F(x_n) + O(\tau^2). \end{aligned}$$

Based on the above two equations, we follow almost the same derivations as [33, Theorem 4.2] to derive the estimate of $e_n^{v_i}$ as

$$\sum_{i=1}^{k} \|e_n^{v_i}\| \le Q\tau,$$

and we invoke this in (4.8) to complete the proof.

5 Numerical Experiments

We carry out a simple numerical experiment to test the convergence rate (denoted by CR) of the scheme (2.3). A detailed comparison between semi-implicit and explicit methods for unconstrained high-index saddle dynamics could be found in [15], which has already indicated the advantages of the semi-implicit method. We apply the Rosenbrock type function

$$E(x_1, x_2, x_3) = a(\sqrt{3}x_2 - 3x_1^2)^2 + b(\sqrt{3}x_1 - 1)^2 + a(\sqrt{3}x_3 - 3x_2^2)^2 + b(\sqrt{3}x_2 - 1)^2.$$

For $(a, b) = (-1, 5.5)$, the point

$$x_* = \mathcal{N}(1,1,1) = \frac{1}{\sqrt{3}}(1,1,1)$$

is an index-1 saddle point of the Rosenbrock type function, while for $(a, b) = (-0.5, 1.5)$, x_* is an index-2 saddle point. We apply the semi-implicit scheme (2.3) to compute the saddle points for these two cases under $T = 10$ and different initial conditions

(a) $x_0 = \mathcal{N}(0.8, 1, 1), \quad v_{1,0} = \mathcal{N}(1, -0.4, -0.4);$

(b) $x_0 = \mathcal{N}(1, 1, 1.4), \quad v_{1,0} = \mathcal{N}(-1, 1, 0);$

(c) $x_0 = \mathcal{N}(0.8, 1, 1), \quad v_{1,0} = \mathcal{N}(1, -0.4, -0.4), \quad v_{2,0} = \mathcal{N}(0, 1, -1);$

(d) $x_0 = \mathcal{N}(1, 1, 1.4), \quad v_{1,0} = \mathcal{N}(-1, 1, 0), \quad v_{2,0} = \mathcal{N}(-0.7, -0.7, 1).$

As the exact trajectory of the constrained high-index saddle dynamics (2.1) is in general not available, we use the numerical solution computed under $\tau = 2^{-13}$ to serve as the reference solution. Numerical results are presented in Tables 1–4, which indicates the first-order accuracy of the semi-implicit scheme (2.3) as proved in Theorem 4.2.

Table 1 CR of computing the index-1 saddle point under the initial condition (a).

τ	$\max_n \|e_n^x\|$	CR	$\max_n \|e_n^{v_1}\|$	CR
2^{-6}	1.65E-02		9.95E-02	
2^{-7}	8.29E-03	0.99	4.47E-02	1.16
2^{-8}	4.09E-03	1.02	2.08E-02	1.10
2^{-9}	1.98E-03	1.04	9.83E-03	1.08

Table 2 CR of computing the index-1 saddle point under the initial condition (b).

τ	$\max_n \|e_n^x\|$	CR	$\max_n \|e_n^{v_1}\|$	CR
2^{-6}	1.03E-02		2.02E-02	
2^{-7}	4.84E-03	1.09	9.53E-03	1.08
2^{-8}	2.32E-03	1.06	4.59E-03	1.05
2^{-9}	1.11E-03	1.06	2.20E-03	1.06

Table 3 CR of computing the index-2 saddle point under the initial condition (c).

τ	$\max_n \|e_n^x\|$	CR	$\max_n \|e_n^{v_1}\|$	CR	$\max_n \|e_n^{v_2}\|$	CR
2^{-6}	1.67E-03		6.06E-02		6.06E-02	
2^{-7}	7.90E-04	1.08	2.87E-02	1.08	2.87E-02	1.08
2^{-8}	3.80E-04	1.05	1.38E-02	1.06	1.38E-02	1.06
2^{-9}	1.82E-04	1.06	6.60E-03	1.06	6.60E-03	1.06

Table 4 CR of computing the index-2 saddle point under the initial condition (d).

τ	$\max_n \|e_n^x\|$	CR	$\max_n \|e_n^{v_1}\|$	CR	$\max_n \|e_n^{v_2}\|$	CR
2^{-6}	2.65E-03		3.53E-02		3.52E-02	
2^{-7}	1.28E-03	1.05	1.69E-02	1.06	1.69E-02	1.06
2^{-8}	6.22E-04	1.04	8.21E-03	1.05	8.18E-03	1.05
2^{-9}	2.99E-04	1.06	3.94E-03	1.06	3.93E-03	1.06

6 Concluding Remarks

In this paper we prove error estimates for the semi-implicit numerical scheme of sphere-constrained high-index saddle dynamics, which ensures the accuracy of performing the saddle dynamics in finding saddle points and constructing the solution landscape for constrained problems. The main difficulties we overcome lie in the semi-implicit treatment on the schemes and the coupling among the dynamics, the retraction, the vector transport and the orthonormalization procedure. Numerical experiments are performed to substantiate the theoretical findings.

There are potential extensions of the current work that deserve further exploration. For instance, the dimer method (see [12]) could be used in (2.1) to approximate the product of the Hessian matrix and the vector for efficient computation and storage, which leads to the shrinking-dimer sphere-constrained high-index saddle dynamics as the unconstrained case (see

[32]). Then the semi-implicit method could be applied to improve the numerical stability that remains to be analyzed.

Furthermore, the ideas and techniques could be employed and improved to analyze the semi-implicit numerical scheme for high-index saddle dynamics constrained by m equalities (see [21, Equation 24]):

$$\begin{cases} \dfrac{\mathrm{d}x}{\mathrm{d}t} = \Big(I - 2\displaystyle\sum_{j=1}^{k} v_j v_j^{\mathrm{T}}\Big) F(x), \\ \dfrac{\mathrm{d}v_i}{\mathrm{d}t} = \Big(I - v_i v_i^{\mathrm{T}} - 2\displaystyle\sum_{j=1}^{i-1} v_j v_j^{\mathrm{T}}\Big) \mathcal{H}(x)[v_i] \\ \qquad\quad -A(x)(A(x)^{\mathrm{T}} A(x))^{-1} \Big(\nabla^2 c(x) \dfrac{\mathrm{d}x}{\mathrm{d}t}\Big)^{\mathrm{T}} v_i, \quad 1 \le i \le k. \end{cases} \tag{6.1}$$

Here $c(x) = (c_1(x), \cdots, c_m(x)) = 0$ represents the m equality constraints and

$$A(x) = (\nabla c_1(x), \cdots, \nabla c_m(x)).$$

The sphere-constrained high-index saddle dynamics (2.1) is a special case of (6.1) with one equality constraint

$$c_1(x) = \|x\| - 1 = 0.$$

In the generalized constrained saddle dynamics (6.1), $\mathcal{H}(x)$ refers to the Riemannian Hessian (see [21]), which is difficult to compute and approximate in practice that we will investigate in the near future.

Declarations

Conflicsts of interest The authors declare no conflicts of interest.

References

[1] Baker, J., An algorithm for the location of transition states, *J. Comput. Chem.*, **7**, 1986, 385–395.

[2] Bao, W. and Cai, Y., Mathematical theory and numerical methods for Bose–Einstein condensation, *Kinet. Relat. Models*, **6**, 2013, 1–135.

[3] Doye, J. and Wales, D., Saddle points and dynamics of Lennard-Jones clusters, solids, and supercooled liquids, *J. Chem. Phys.*, **116**, 2002, 3777–3788.

[4] E, W. and Vanden-Eijnden, E., Transition-path theory and path-finding algorithms for the study of rare events, *Annu. Rev. Phys. Chem.*, **61**, 2010, 391–420.

[5] E, W. and Zhou, X., The gentlest ascent dynamics, *Nonlinearity*, **24**, 2011, 1831–1842.

[6] Farrell, P., Birkisson, Á. and Funke, S., Deflation techniques for finding distinct solutions of nonlinear partial differential equations, *SIAM J. Sci. Comput.*, **37**, 2015, A2026–A2045.

[7] Gao, W., Leng, J. and Zhou, X., An iterative minimization formulation for saddle point search, *SIAM J. Numer. Anal.*, **53**, 2015, 1786–1805.

[8] Gould, N., Ortner, C. and Packwood, D., A dimer-type saddle search algorithm with preconditioning and linesearch, *Math. Comp.*, **85**, 2016, 2939–2966.

[9] Grantham, W., Gradient transformation trajectory following algorithms for determining stationary min-max saddle points, in Advances in Dynamic Game Theory, Ann. Internat. Soc. Dynam. Games, **9**, Birkhauser Boston, Boston, MA, 2007, 639–657.

[10] Han, Y., Hu, Y., Zhang, P., et al., Transition pathways between defect patterns in confined nematic liquid crystals, *J. Comput. Phys.*, **396**, 2019, 1–11.

[11] Han, Y., Xu, Z., Shi, A. and Zhang, L., Pathways connecting two opposed bilayers with a fusion pore: A molecularly-informed phase field approach, *Soft Matter*, **16**, 2020, 366–374.

[12] Henkelman, G. and Jónsson, H., A dimer method for finding saddle points on high dimensional potential surfaces using only first derivatives, *J. Chem. Phys.*, **111**, 1999, 7010–7022.

[13] Levitt, A. and Ortner, C., Convergence and cycling in walker-type saddle search algorithms, *SIAM J. Numer. Anal.*, **55**, 2017, 2204–2227.

[14] Li, Y. and Zhou, J., A minimax method for finding multiple critical points and its applications to semilinear PDEs, *SIAM J. Sci. Comput.*, **23**, 2001, 840–865.

[15] Luo, Y., Zhang, L., Zhang, P., et al., Semi-implicit method of high-index saddle dynamics and application to construct solution landscape, Submitted.

[16] Mehta, D., Finding all the stationary points of a potential-energy landscape via numerical polynomial-homotopy-continuation method, *Phys. Rev. E*, **84**, 2011, 025702.

[17] Milnor, J., Morse Theory, Ann. of Math. Stud., **51**, Princeton University Press, NJ, 1963.

[18] Thomson, J., XXIV, On the structure of the atom: An investigation of the stability and periods of oscillation of a number of corpuscles arranged at equal intervals around the circumference of a circle; with application of the results to the theory of atomic structure, *London, Edinburgh, Dublin Phil. Mag. J. Sci.*, **7**, 1904, 237–265.

[19] Wang, W., Zhang, L. and Zhang, P., Modelling and computation of liquid crystals, *Acta Numerica*, **30**, 2021, 765–851.

[20] Xie, Z., Yuan, Y. and Zhou, J., On solving semilinear singularly perturbed Neumann problems for multiple solutions, *SIAM J. Sci. Comput.*, **44**, 2022, A501–A523.

[21] Yin, J., Huang, Z. and Zhang, L., Constrained high-index saddle dynamics for the solution landscape with equality constraints, *J. Sci. Comput.*, **91**, 2022, 62.

[22] Yin, J., Jiang, K., Shi, A., et al., Transition pathways connecting crystals and quasicrystals, *Proc. Natl. Acad. Sci. U.S.A.*, **118**, 2021, e2106230118.

[23] Yin, J., Wang, Y., Chen, J., et al., Construction of a pathway map on a complicated energy landscape, *Phys. Rev. Lett.*, **124**, 2020, 090601.

[24] Yin, J., Yu, B. and Zhang, L., Searching the solution landscape by generalized high-index saddle dynamics, *Sci. China Math.*, **64**, 2021, 1801.

[25] Yin, J., Zhang, L. and Zhang, P., High-index optimization-based shrinking dimer method for finding high-index saddle points, *SIAM J. Sci. Comput.*, **41**, 2019, A3576–A3595.

[26] Yu, B., Zheng, X., Zhang, P. and Zhang, L., Computing solution landscape of nonlinear space-fractional problems via fast approximation algorithm, *J. Comput. Phys.*, **468**, 2022, 111513.

[27] Zhang, L., Chen, L. and Du, Q., Morphology of critical nuclei in solid-state phase transformations, *Phys. Rev. Lett.*, **98**, 2007, 265703.

[28] Zhang, L., Chen, L. and Du, Q., Simultaneous prediction of morphologies of a critical nucleus and an equilibrium precipitate in solids, *Commun. Comput. Phys.*, **7**, 2010, 674–682.

[29] Zhang, J. and Du, Q., Shrinking dimer dynamics and its applications to saddle point search, *SIAM J. Numer. Anal.*, **50**, 2012, 1899–1921.

[30] Zhang, L., Ren, W., Samanta, A. and Du, Q., Recent developments in computational modelling of nucleation in phase transformations, *npj Comput. Mater.*, **2**, 2016, 16003.

[31] Zhang, L., Zhang, P. and Zheng, X., Error estimates of Euler discretization to high-index saddle dynamics, *SIAM J. Numer. Anal.*, **60**, 2022, 2925–2944.

[32] Zhang, L., Zhang, P. and Zheng, X., Mathematical and numerical analysis to shrinking-dimer saddle dynamics with local Lipschitz conditions, *CSIAM Trans. Appl. Math.*, **4**, 2023, 157–176.

[33] Zhang, L., Zhang, P. and Zheng, X., Discretization and index-robust error analysis for constrained high-index saddle dynamics on high-dimensional sphere, *Sci. China Math.*, 2023, DOI: 10.1007/s11425-022-2149-2.