

Contraction Metrics by Numerical Integration and Quadrature: Uniform Error Estimate

Peter Giesl¹ ^a, Sigurdur Hafstein² ^b and Iman Mehrabinezhad² ^c

¹*Department of Mathematics, University of Sussex, Falmer, BN1 9QH, U.K.*

²*Science Institute, University of Iceland, Dunhagi 3, 107 Reykjavík, Iceland*

Keywords: Contraction Metric, Numerical Method, Error Estimate.

Abstract: We show that contraction metrics for continuous time dynamical systems can be computed numerically using numerical integration of certain initial value problems with a subsequent numerical quadrature. Further, we show that for any compact subset of an equilibrium's basin of attraction and any $\epsilon > 0$, the parameters for the numerical methods, i.e. the integration interval and the step-size, can be chosen such that the error in the contraction metric is less than ϵ at any point in the compact subset. These results will be used as a part of a numerical method to rigorously compute contraction metrics.

1 INTRODUCTION

We consider the system

$$\dot{\mathbf{x}} = \mathbf{f}(\mathbf{x}), \quad \mathbf{f} \in C^s(\mathbb{R}^n; \mathbb{R}^n), \quad s \geq 1. \quad (1)$$


The solution $\mathbf{x}(t)$ to the initial value problem (1) with $\mathbf{x}(0) = \boldsymbol{\xi}$ is denoted by $\boldsymbol{\phi}(t, \boldsymbol{\xi})$. A contraction metric for system (1) is a Riemannian metric defined on a positively invariant set of the dynamics, such that the distance between adjacent trajectories is decreasing with respect to the metric. The existence of a contraction metric asserts the existence of exactly one equilibrium point inside a positively invariant and connected set and that it is exponentially stable.


Contraction metrics have received considerable attention in the literature (Lewis, 1949; Lewis, 1951; Demidovič, 1961; Krasovskiĭ, 1963; Borg, 1960; Hartman, 1961; Hartman, 1964; Lohmiller and Slotine, 1998; Aminzare and Sontag, 2014; Simpson-Porco and Bullo, 2014; Forni and Sepulchre, 2014; Giesl, 2015), as they can characterize the long term behavior of system (1). Since many phenomena in engineering and science are modelled by system (1), contraction metrics are of much value in understanding real-world systems.


Since the analytical computation of a contraction metric for a nonlinear system is notoriously diffi-

cult, numerical methods have been considered (Aylward et al., 2008; Giesl and Hafstein, 2013; Giesl, 2019; Giesl et al., 2023a), see also the recent review (Giesl et al., 2023b). To advance such methods we present a novel theorem that shows that contraction metrics can be approximated arbitrarily close to the analytic solution, uniformly on any compact subset K of an exponentially stable equilibrium's basin of attraction, using numerical integration and quadrature. These results are essential in developing combined approximation-verification methods to rigorously compute contraction metrics, as in (Giesl et al., 2021a; Giesl et al., 2021b), but using numerical integration and quadrature for the approximation instead of generalized interpolation in reproducing kernel Hilbert spaces (Giesl et al., 2023c).

Let us give an overview of the paper: In Section 2, we recall some facts about contraction metrics, including an existence result of a contraction metric given by an integral formula. In Section 3, we numerically approximate a contraction metric using numerical integration, with the fourth-order Adams-Bashforth (AB4) multi-step scheme initialized with fourth-order Runge-Kutta (RK4), and subsequently we use numerical quadrature to approximately integrate the results; we also derive error bounds for these methods. These estimates are then used to prove the main result of the paper, Theorem 4.1 presented in Section 4, before we conclude our work in Section 5. **Notation:** We write $\mathbb{N}_0 := \{0, 1, 2, \dots\}$ for the natural numbers, including zero, and $\mathbb{N}_+ := \mathbb{N}_0 \setminus \{0\}$ for

^a  <https://orcid.org/0000-0003-1421-6980>

^b  <https://orcid.org/0000-0003-0073-2765>

^c  <https://orcid.org/0000-0002-6346-9901>

the positive natural numbers. We denote the usual p -norms on \mathbb{R}^n and the corresponding induced matrix norms by $\|\cdot\|_p$, $1 \leq p < \infty$. For both vectors in \mathbb{R}^n and matrices in $\mathbb{R}^{n \times n}$ we write $\|\cdot\|_{\max}$ for the maximum absolute value norm, i.e. $\|\mathbf{x}\|_{\max} := \max_{i=1,2,\dots,n} |x_i|$ for a vector $\mathbf{x} \in \mathbb{R}^{n \times n}$ and $\|A\|_{\max} := \max_{i,j=1,2,\dots,n} |a_{ij}|$ for a matrix $A = (a_{ij}) \in \mathbb{R}^{n \times n}$. Apart from the usual equivalence estimates for the p -norms on \mathbb{R}^n , recall the norm equivalence $\|A\|_{\max} \leq \|A\|_2 \leq n\|A\|_{\max}$ for a matrix $A \in \mathbb{R}^{n \times n}$ and that $\|\cdot\|_{\max}$ is not sub-multiplicative, but $\|Ab\|_{\max} \leq n\|A\|_{\max}\|b\|_{\max}$ for $b := B \in \mathbb{R}^{n \times n}$ or $b := \mathbf{b} \in \mathbb{R}^n$. We denote the symmetric $n \times n$ matrices with real entries by $\mathbb{S}^{n \times n}$ and we write I for the $n \times n$ identity matrix (n can always be determined from the context). $A \preceq B$ for $A, B \in \mathbb{S}^{n \times n}$ means that the matrix $A - B$ is negative semi-definite, i.e. $\mathbf{x}^T(A - B)\mathbf{x} \leq 0$ for all $\mathbf{x} \in \mathbb{R}^n$.

2 CONTRACTION METRICS

We first review basic concepts about Riemannian contraction metrics that are used in this paper.

Definition 2.1. (Riemannian metric, contraction metric) Let K be a compact subset of an open set $G \subset \mathbb{R}^n$. A function $M \in C^1(G; \mathbb{S}^{n \times n})$ is called a Riemannian metric if $M(\mathbf{x})$ is positive definite at every $\mathbf{x} \in G$. The Riemannian metric M is said to be a contraction metric for system (1), contracting on K , if

$$M(\mathbf{x})D\mathbf{f}(\mathbf{x}) + D\mathbf{f}(\mathbf{x})^T M(\mathbf{x}) + M'(\mathbf{x}) \preceq -2\nu M(\mathbf{x}),$$

for some $\nu > 0$ and all $\mathbf{x} \in K$.

In Definition 2.1

$$\begin{aligned} M'(\mathbf{x}) &:= \left. \frac{d}{dt} M(\boldsymbol{\phi}(t, \mathbf{x})) \right|_{t=0} \\ &= \left(\nabla M_{ij}(\mathbf{x}) \cdot \mathbf{f}(\mathbf{x}) \right)_{i,j \in \{1,2,\dots,n\}} \end{aligned}$$

is the orbital derivative of M along the solutions of (1).

The following theorem follows directly from (Giesl and Wendland, 2019, Thms. 2.2 and 2.3). In the formula (2), $\tau \mapsto \boldsymbol{\phi}(\tau, \mathbf{x})$ is the solution to (1) with initial value $\boldsymbol{\phi}(0, \mathbf{x}) = \mathbf{x}$ and $\tau \mapsto \boldsymbol{\psi}(\tau, \mathbf{x})$ is the matrix-valued solution to

$$\dot{Y} = D\mathbf{f}(\boldsymbol{\phi}(t, \mathbf{x}))Y, \quad Y(0) = I,$$

i.e. $\tau \mapsto \boldsymbol{\psi}(\tau, \mathbf{x})$ is the principal fundamental matrix solution.

Theorem 2.2. (Existence of a contraction metric) Let $\mathbf{f} \in C^s(\mathbb{R}^n; \mathbb{R}^n)$, $s \geq 2$. Let \mathbf{x}_0 be an exponentially stable equilibrium of (1) with basin of attraction $\mathcal{A}(\mathbf{x}_0) := \{\mathbf{x} \in \mathbb{R}^n : \lim_{t \rightarrow \infty} \boldsymbol{\phi}(t, \mathbf{x}) = \mathbf{x}_0\}$. Let

$C \in C^{s-1}(\mathcal{A}(\mathbf{x}_0); \mathbb{S}^{n \times n})$ be such that $C(\mathbf{x})$ is a positive definite matrix for all $\mathbf{x} \in \mathcal{A}(\mathbf{x}_0)$. Then $M \in C^{s-1}(\mathcal{A}(\mathbf{x}_0); \mathbb{S}^{n \times n})$, given by the formula

$$M(\boldsymbol{\xi}) = \int_0^\infty \boldsymbol{\psi}(\tau, \boldsymbol{\xi})^T C(\boldsymbol{\phi}(\tau, \boldsymbol{\xi})) \boldsymbol{\psi}(\tau, \boldsymbol{\xi}) d\tau, \quad (2)$$

is a contraction metric for (1), that is contracting on any compact $K \subset \mathcal{A}(\mathbf{x}_0)$.

In the following section we will show that the contraction metric $M(\boldsymbol{\xi})$ in formula (2) can be estimated arbitrarily close by using numerical integration and numerical quadrature.

3 ESTIMATION METHOD

In this section we describe in detail how we estimate $M(\boldsymbol{\xi})$ in formula (2) by $\tilde{M}(\boldsymbol{\xi})$ at a point $\boldsymbol{\xi} \in \mathbb{R}^n$ in three steps and we conclude with an error estimate in Theorem 4.1. We first fix a matrix-valued function $C \in C^{s-1}(\mathbb{R}^n; \mathbb{S}^{n \times n})$, which in practice can be taken simply as the constant identity matrix $I \in \mathbb{R}^{n \times n}$, a time-horizon $H > 0$, and a set of points X , at which we compute values for our metric \tilde{M} inspired by (2). For $\boldsymbol{\xi} \in X$ we first compute a numerical approximation $\tilde{\boldsymbol{\phi}} : [0, H] \rightarrow \mathbb{R}^n$ to the initial-value problem

$$\dot{\mathbf{x}} = \mathbf{f}(\mathbf{x}), \quad \mathbf{x}(0) = \boldsymbol{\xi}, \quad (3)$$

on the time-horizon $[0, H]$. We do this by fixing the number of time-steps N and the corresponding length of a uniform time-step $h := H/N$ and then generate a sequence of vectors $\tilde{\boldsymbol{\phi}}_i$, $i = 0, 1, \dots, N$, such that $\tilde{\boldsymbol{\phi}}_i$ approximates the true solution $\boldsymbol{\phi}(\cdot, \boldsymbol{\xi})$ to the initial-value problem (3) at time $t_i := ih$, i.e.

$$\tilde{\boldsymbol{\phi}}_i \approx \boldsymbol{\phi}(t_i, \boldsymbol{\xi}).$$

In the sequel, when we refer to $\tilde{\boldsymbol{\phi}} = \tilde{\boldsymbol{\phi}}(t, \boldsymbol{\xi})$ as a function, we mean the linear interpolation of the values $\tilde{\boldsymbol{\phi}}_i$, i.e.

$$\tilde{\boldsymbol{\phi}}(t) = \frac{t - t_i}{t_{i+1} - t_i} (\tilde{\boldsymbol{\phi}}_{i+1} - \tilde{\boldsymbol{\phi}}_i) + \tilde{\boldsymbol{\phi}}_i \quad \text{when } t_i \leq t \leq t_{i+1}.$$

Then we use our approximate solution $\tilde{\boldsymbol{\phi}}$ to (3) to obtain an approximation \tilde{Y} of the principal fundamental matrix solution to $\dot{Y} = D\mathbf{f}(\boldsymbol{\phi}(t, \boldsymbol{\xi}))Y$. That is, we solve numerically the matrix-valued initial-value problem

$$\dot{Y} = g(t)Y, \quad Y(0) = I, \quad (4)$$

where $D\mathbf{f}(\boldsymbol{\phi}(t, \boldsymbol{\xi}))$ has been substituted by the approximation $g(t) := D\mathbf{f}(\tilde{\boldsymbol{\phi}}(t, \boldsymbol{\xi}))$. Finally, we use our numerical solutions $\tilde{\boldsymbol{\phi}}$ to (3) and \tilde{Y} to (4) to compute an approximation

$$\tilde{M}(\boldsymbol{\xi}) \approx \int_0^H Y(\tau)^T C(\boldsymbol{\phi}(\tau, \boldsymbol{\xi})) Y(\tau) d\tau \quad (5)$$

to $M(\xi)$ using a Romberg-like numerical quadrature. Note that there are several approximations to the actual integral: firstly, we use a Romberg-like numerical quadrature to compute the integral, for which we now only need values of the integrand at discrete time steps; secondly, we replace the values of the integrand with our numerical solutions $\tilde{\phi}_i$ and \tilde{Y}_i to (3) and (4), respectively.

For the initial value problems (3) and (4) we use the Adams-Bashforth method of order 4 (AB4) initialized with the usual Runge-Kutta method of order 4 (RK4). The formula for a general initial-value problem of the form

$$\dot{\mathbf{z}} = \mathbf{v}(t, \mathbf{z}), \quad \mathbf{z}(t_0) = \xi, \quad (6)$$

to generate the approximations $\tilde{\mathbf{z}}_i \approx \mathbf{z}(t_i, \xi)$, where $t_i = ih + t_0$ and $\mathbf{z}(t_i, \xi)$ is the value of the true solution $t \mapsto \mathbf{z}(t, \xi)$ to (6) at time t_i , is

$$\tilde{\mathbf{z}}_{i+1} = \tilde{\mathbf{z}}_i + \frac{h}{24}(55\mathbf{v}_i - 59\mathbf{v}_{i-1} + 37\mathbf{v}_{i-2} - 9\mathbf{v}_{i-3}), \quad (7)$$

for AB4, where $\mathbf{v}_i := \mathbf{v}(t_i, \tilde{\mathbf{z}}_i)$. Since AB4 is a multi-step method we use RK4 to compute the first 3 steps $\tilde{\mathbf{z}}_1, \tilde{\mathbf{z}}_2, \tilde{\mathbf{z}}_3$ after $\tilde{\mathbf{z}}_0 := \xi$, needed to initialize it. In more detail, for the initialization we set

$$\mathbf{k}_1 = h\mathbf{v}(t_i, \tilde{\mathbf{z}}_i) \quad (8)$$

$$\mathbf{k}_2 = h\mathbf{v}(t_i + h/2, \tilde{\mathbf{z}}_i + \mathbf{k}_1/2)$$

$$\mathbf{k}_3 = h\mathbf{v}(t_i + h/2, \tilde{\mathbf{z}}_i + \mathbf{k}_2/2)$$

$$\mathbf{k}_4 = h\mathbf{v}(t_i + h, \tilde{\mathbf{z}}_i + \mathbf{k}_3)$$

$$\tilde{\mathbf{z}}_{i+1} = \tilde{\mathbf{z}}_i + \frac{1}{6}(\mathbf{k}_1 + 2\mathbf{k}_2 + 2\mathbf{k}_3 + \mathbf{k}_4)$$

for $i = 0, 1, 2$.

In the following, we fix a compact set $S \subset \mathbb{R}^n$ which is positively invariant, both for the solution ϕ and its numerical approximation sequences $\tilde{\phi}_i$. Hence, the values $\tilde{\phi}(t)$, $t \in [0, H]$, are in the convex hull of S , independent of the initial value $\xi \in S$. Note that we have a Lipschitz constant for any continuously differentiable functions on S or its (compact) convex hull. Level sets of numerically computed Lyapunov-like functions can be used for identifying such sets, see (Giesl et al., 2023d, Thm. 3.4). In (Giesl et al., 2023d, Thms. 2.1 and 3.5) it is established that for a given compact set $K \subset \mathcal{A}(\mathbf{x}_0)$ there exists a compact set $S \supset K$, which is positively invariant for both the solution and its numerical approximation by our numerical method, i.e. AB4 initialized with RK4, if the time-steps h are sufficiently small, namely $h \leq h'$, where $h' > 0$ is a constant depending on K and \mathbf{f} . Subsequently we additionally assume that h is smaller than other constants for additional estimates to hold true.

In the following subsections we explain and estimate the errors of the numerical approximations to the solution $\phi(t, \xi)$, the solution of $\dot{Y} = D\mathbf{f}(\phi(t, \xi))Y$ and finally the numerical quadrature of the integral in preparation for the main result, Theorem 4.1.

3.1 Step I: Numerical Approximation of $\dot{\mathbf{x}} = \mathbf{f}(\mathbf{x})$

For the initial value problem (3) we use AB4 initialized with RK4; i.e. we set $\mathbf{v}(t, \mathbf{z}) = \mathbf{f}(\mathbf{z})$ in (6). The true solution $\mathbf{z}(t, \xi)$ is denoted by $\phi(t, \xi)$ and its approximations at $t_i = hi$ by $\tilde{\phi}_i$. Thus, the formulas (8) become

$$\mathbf{k}_1 = h\mathbf{f}(\tilde{\phi}_i) \quad (9)$$

$$\mathbf{k}_2 = h\mathbf{f}(\tilde{\phi}_i + \mathbf{k}_1/2)$$

$$\mathbf{k}_3 = h\mathbf{f}(\tilde{\phi}_i + \mathbf{k}_2/2)$$

$$\mathbf{k}_4 = h\mathbf{f}(\tilde{\phi}_i + \mathbf{k}_3)$$

$$\tilde{\phi}_{i+1} = \tilde{\phi}_i + \frac{1}{6}(\mathbf{k}_1 + 2\mathbf{k}_2 + 2\mathbf{k}_3 + \mathbf{k}_4),$$

where $i = 0, 1, 2$. And for $i \geq 3$ we use AB4 and (7) becomes

$$\tilde{\phi}_{i+1} = \tilde{\phi}_i + \frac{h}{24}(55\mathbf{f}(\tilde{\phi}_i) - 59\mathbf{f}(\tilde{\phi}_{i-1}) + 37\mathbf{f}(\tilde{\phi}_{i-2}) - 9\mathbf{f}(\tilde{\phi}_{i-3})). \quad (10)$$

In the following we assume that a fixed time-horizon $H > 0$ is given such that $t_i \leq H$ for all i . The error of approximation in the initializing step using RK4 is bounded by $C_1 h^5$, and the rest of the $\tilde{\phi}_i$ sequence using AB4 is bounded by the local error $C_2 h^5$, where C_1 and C_2 are constants that depend on the fourth order derivative of \mathbf{f} in S , which is positively invariant both for the true solution and the approximation, see (Giesl et al., 2023d, Thm. 3.5). We recall below the well known results, that then the global error is bounded by $C_4 h^4$ on the interval $[0, H]$.

Since we initialized the AB4 method using RK4 and not the exact values of the corresponding right-hand side function \mathbf{f} , there will be a small error accumulating through the algorithm. In more detail, let $L > 0$ be a Lipschitz constant for \mathbf{f} on $S \subset \mathbb{R}^n$. If \mathbf{y} and \mathbf{z} are solutions in S with initial conditions $\mathbf{y}(a)$ and $\mathbf{z}(a)$ at time $t = a$ respectively, then it is well known that Gronwall's inequality delivers

$$\|\mathbf{y}(t) - \mathbf{z}(t)\|_{\max} \leq e^{L|t-a|} \|\mathbf{y}(a) - \mathbf{z}(a)\|_{\max} \quad (11)$$

as long as the solutions stay in S .

Let us recall how the local errors accumulate to form global errors in the multi-step scenario, see e.g. (Sauer, 2012; Deuffhard and Hohmann, 2008),

because we will use similar reasoning in the following. At the initial condition $\tilde{\mathbf{x}}_0 = \boldsymbol{\xi}$, the global error is $g_0 = \|\tilde{\mathbf{x}}_0 - \mathbf{x}_0\|_{\max} = \|\boldsymbol{\xi} - \boldsymbol{\xi}\|_{\max} = 0$. After one step, there is no accumulated error from previous steps, and the bound on the global error is the local truncation error, $g_1 = e_1 = \|\tilde{\mathbf{x}}_1 - \mathbf{x}_1\|_{\max} \leq C_1 h^5$. After two steps, we break g_2 down into the local truncation error plus the accumulated error from the earlier step. Define $\mathbf{z}(t)$ to be the solution of the initial value problem

$$\begin{cases} \dot{\mathbf{x}} = \mathbf{f}(\mathbf{x}) \\ \mathbf{x}(t_1) = \tilde{\mathbf{x}}_1, \\ t \in [t_1, t_2] \end{cases} \quad (12)$$

Thus, $\mathbf{z}(t_2)$ is the exact value of the solution at $t = t_2$ starting at initial condition $(t_1, \tilde{\mathbf{x}}_1)$. Note that if we used the initial condition (t_1, \mathbf{x}_1) , we would get \mathbf{x}_2 , which is on the actual solution curve, unlike $\mathbf{z}(t_2)$. Then $e_2 = \|\tilde{\mathbf{x}}_2 - \mathbf{z}(t_2)\|_{\max} \leq C_1 h^5$ is the local truncation error of step $i = 2$. The other difference $\|\mathbf{z}(t_2) - \mathbf{x}_2\|_{\max}$ is covered by equation (11), since it is the difference between two solutions of the same equation with different initial conditions $\tilde{\mathbf{x}}_1$ and \mathbf{x}_1 . Therefore,

$$\begin{aligned} g_2 &= \|\tilde{\mathbf{x}}_2 - \mathbf{x}_2\|_{\max} \\ &\leq \|\tilde{\mathbf{x}}_2 - \mathbf{z}(t_2)\|_{\max} + \|\mathbf{z}(t_2) - \mathbf{x}_2\|_{\max} \\ &\leq e_2 + e^{Lh} g_1 \\ &= e_2 + e^{Lh} e_1 \\ &\leq C_1 h^5 (1 + e^{Lh}). \end{aligned}$$

The argument is the same for step $i = 3$, which yields

$$\begin{aligned} g_3 &= \|\tilde{\mathbf{x}}_3 - \mathbf{x}_3\|_{\max} \leq e_3 + e^{Lh} g_2 \\ &\leq e_3 + e^{Lh} e_2 + e^{2Lh} e_1 \leq C_1 h^5 (1 + e^{Lh} + e^{2Lh}). \end{aligned}$$

For step $i = 4$, the initializing phase with RK4 is finished and we have a different local truncation error term, namely $e_4 = \|\tilde{\mathbf{x}}_4 - \mathbf{z}(t_4)\|_{\max} \leq C_2 h^5$. Thus,

$$\begin{aligned} g_4 &= \|\tilde{\mathbf{x}}_4 - \mathbf{x}_4\|_{\max} \\ &\leq \|\tilde{\mathbf{x}}_4 - \mathbf{z}(t_4)\|_{\max} + \|\mathbf{z}(t_4) - \mathbf{x}_4\|_{\max} \\ &\leq e_4 + e^{Lh} g_3 \\ &\leq C_2 h^5 + e^{Lh} C_1 h^5 (1 + e^{Lh} + e^{2Lh}) \\ &\leq C_3 h^5 (1 + e^{Lh} + e^{2Lh} + e^{3Lh}), \end{aligned}$$

where we have introduced the new constant $C_3 := \max(C_1, C_2) > 0$. Now the rest can be done similarly

to get the global truncation error at any step i .

$$\begin{aligned} g_i &= \|\tilde{\mathbf{x}}_i - \mathbf{x}_i\|_{\max} \\ &\leq e_i + e^{Lh} e_{i-1} + e^{2Lh} e_{i-2} + \dots + e^{(i-1)Lh} e_1 \\ &\leq C_3 h^5 (1 + e^{Lh} + \dots + e^{(i-1)Lh}) \\ &= C_3 h^5 \frac{e^{iLh} - 1}{e^{Lh} - 1} \leq \frac{C_3 h^4}{L} (e^{Li} - 1) \\ &\leq \frac{C_3}{L} (e^{LH} - 1) h^4 =: C_4 h^4, \end{aligned} \quad (13)$$

where C_4 depends on H . Applying this result to $\mathbf{x}_i = \boldsymbol{\phi}(t_i, \boldsymbol{\xi})$ and $\tilde{\mathbf{x}}_i = \tilde{\boldsymbol{\phi}}_i$ gives us the desired estimate.

3.2 Step II: Numerical Approximation of $\dot{Y} = G(t)Y$

Once more, we use AB4 and RK4 to solve the initial value problem (4) numerically. We are still assuming that the time-horizon $H > 0$ from Step I is fixed. Note that with $\mathbf{y}^1(t), \mathbf{y}^2(t), \dots, \mathbf{y}^n(t)$ as the column vectors of the matrix $Y = Y(t)$ in (4), i.e.

$$Y(t) = \begin{pmatrix} | & | & & | \\ \mathbf{y}^1(t) & \mathbf{y}^2(t) & \dots & \mathbf{y}^n(t) \\ | & | & & | \end{pmatrix},$$

the matrix-valued initial-value problem (4) boils down to the vector-valued initial-value problems

$$\dot{\mathbf{y}}^j = g(t)\mathbf{y}^j, \quad \mathbf{y}^j(0) = \mathbf{e}_j, \quad j = 1, 2, \dots, n, \quad (14)$$

where \mathbf{e}_j is the usual j th unit vector in \mathbb{R}^n .

Note that for $\mathbf{y}(t) = \mathbf{y}^j(t)$ with any $j = 1, \dots, n$ and $g(t) = D\mathbf{f}(\boldsymbol{\phi}(t, \boldsymbol{\xi}))$ we have with $M = \max_{\mathbf{x} \in S} \|D\mathbf{f}(\mathbf{x})\|_2$ in the positively invariant set S

$$\begin{aligned} \|\mathbf{y}(t)\|_{\max} &\leq \|\mathbf{y}(t)\|_2 \\ &\leq \|\mathbf{y}(0)\|_2 + \int_0^t \|D\mathbf{f}(\boldsymbol{\phi}(s, \boldsymbol{\xi}))\|_2 \|\mathbf{y}(s)\|_2 ds \\ &\leq \|\mathbf{y}(0)\|_2 + \int_0^t M \|\mathbf{y}(s)\|_2 ds \\ &\leq \|\mathbf{y}(0)\|_2 \exp(tM) \\ &\leq n \|\mathbf{y}(0)\|_{\max} \exp(tM), \end{aligned}$$

where we have used Gronwall's lemma (Walter, 1998). Hence, there exists a constant such that $\|Y(t)\| \leq C$ holds for all $t \in [0, H]$; note that C depends on H . This also implies that the derivatives with respect to (t, \mathbf{y}) up to order 4 of the right-hand side, namely $D\mathbf{f}(\boldsymbol{\phi}(t, \boldsymbol{\xi}))\mathbf{y}$ are bounded, if $\mathbf{f} \in C^5$, uniformly for $\boldsymbol{\xi} \in S$, and the approximation using RK4 and AB4 is bounded by a constant times h^4 . We will assume that, in addition to the previous assumptions $h \leq h'$ that h is also bounded by the constant $\min(h^*, h^{**}, 1)$, which depends on H and S , see (22) and (25).

However, in our computations we need to replace $\phi(\cdot, \xi)$ by $\tilde{\phi} \approx \phi(\cdot, \xi)$. To study the error, we will altogether consider and compare three approximate solutions of $\dot{Y} = Df(\phi(t, \xi))Y$, $Y(0) = I$:

1. We denote by $Y(t)$ the solution of $\dot{Y} = Df(\phi(t, \xi))Y$, $Y(0) = I$ at a given time t . The values at times t_i are denoted by $Y_i = Y(t_i)$.
2. By \tilde{Y}_{ϕ_i} we denote the numerical approximation of $Y(t)$ at time t_i using RK4 and AB4 with the true solution ϕ in the formulas; these formulas use k_1, \dots, k_4 , see (15) and (16).
3. Finally, we denote by \tilde{Y}_i the approximation of $Y(t)$, using $\tilde{\phi}$ in the RK4 and AB4 formula; these formulas use k'_1, \dots, k'_4 , see (17) and (18).

The RK4 formulas (8) for the values \tilde{Y}_{ϕ_i} with the correct solution ϕ are given by

$$\begin{aligned} k_1 &= hDf(\phi_i)\tilde{Y}_{\phi_i} \\ k_2 &= hDf(\phi_{i+\frac{1}{2}})(\tilde{Y}_{\phi_i} + k_1/2) \\ k_3 &= hDf(\phi_{i+\frac{1}{2}})(\tilde{Y}_{\phi_i} + k_2/2) \\ k_4 &= hDf(\phi_{i+1})(\tilde{Y}_{\phi_i} + k_3) \end{aligned} \quad (15)$$

$$\tilde{Y}_{\phi_{i+1}} = \tilde{Y}_{\phi_i} + \frac{1}{6}(k_1 + 2k_2 + 2k_3 + k_4).$$

Note that we need $\phi(t_i + h/2, \xi)$, which we denoted $\phi_{i+\frac{1}{2}}$ in the formulas. The AB4 method for (4) is given by

$$\begin{aligned} \tilde{Y}_{\phi_{i+1}} &= \tilde{Y}_{\phi_i} + \frac{h}{24} [55Df(\phi_i)\tilde{Y}_{\phi_i} - 59Df(\phi_{i-1})\tilde{Y}_{\phi_{i-1}} \\ &\quad + 37Df(\phi_{i-2})\tilde{Y}_{\phi_{i-2}} - 9Df(\phi_{i-3})\tilde{Y}_{\phi_{i-3}}], \end{aligned} \quad (16)$$

and we do not need $\phi_{i+\frac{1}{2}}$ for $i \geq 3$.

For our actual computation of the values \tilde{Y}_i , we use estimated values $\tilde{\phi}_i$ of $\phi(t_i, \xi)$. Hence, the RK4 formulas are given by

$$\begin{aligned} k'_1 &= hDf(\tilde{\phi}_i)\tilde{Y}_i \\ k'_2 &= hDf(\tilde{\phi}_{i+\frac{1}{2}})(\tilde{Y}_i + k'_1/2) \\ k'_3 &= hDf(\tilde{\phi}_{i+\frac{1}{2}})(\tilde{Y}_i + k'_2/2) \\ k'_4 &= hDf(\tilde{\phi}_{i+1})(\tilde{Y}_i + k'_3) \end{aligned} \quad (17)$$

$$\tilde{Y}_{i+1} = \tilde{Y}_i + \frac{1}{6}(k'_1 + 2k'_2 + 2k'_3 + k'_4).$$

Note that we need estimates $\tilde{\phi}_{i+\frac{1}{2}} \approx \phi(t_i + h/2, \xi)$ in the formulas. Thus, we need to use RK4 with time-steps $h_0 := h/2$ for the initial-value problem (3), when we want to use its results in the RK4 formula (17)

for the initial-value problem (4) to compute the values $\tilde{Y}_1, \tilde{Y}_2, \tilde{Y}_3$. The AB4 method is given by

$$\begin{aligned} \tilde{Y}_{i+1} &= \tilde{Y}_i + \frac{h}{24} [55Df(\tilde{\phi}_i)\tilde{Y}_i - 59Df(\tilde{\phi}_{i-1})\tilde{Y}_{i-1} \\ &\quad + 37Df(\tilde{\phi}_{i-2})\tilde{Y}_{i-2} - 9Df(\tilde{\phi}_{i-3})\tilde{Y}_{i-3}], \end{aligned} \quad (18)$$

and we do not need $\tilde{\phi}_{i+\frac{1}{2}}$ for $i \geq 3$.

Similar to the arguments of the previous subsection, we will show that the global error of approximation $\|Y_i - \tilde{Y}_i\|_{\max}$ is bounded by $C_5 h^4$ for a constant $C_5 > 0$. In this case, however, there are two sources for the error; one is using the RK4 and AB4 methods to solve the differential equation (4) numerically and the other comes from the fact that we use the approximations $\tilde{\phi}_i$ instead of the correct values $\phi(t_i, \xi)$.

For the difference between Y_i and \tilde{Y}_{ϕ_i} , we will use the global error estimates of RK4 and AB4, while for the difference between \tilde{Y}_{ϕ_i} and \tilde{Y}_i , we estimate using the formulas directly.

We start with the first task and note that a similar result to (13) holds also in the case of time-dependent right-hand sides to conclude that

$$\|Y_i - \tilde{Y}_{\phi_i}\|_{\max} \leq C_5 h^4 \quad (19)$$

for $i = 1, \dots, N$, where $Nh = H$ and the constant C_5 depends on H and the derivatives of up to order 5 of f in S . Note that the derivatives up to order 4 of $v(t, z) = Df(\phi(t, \xi))z$ for $\xi \in S$ exist and are uniformly bounded by a constant; see the argumentation at the beginning of the section. This implies that there is a constant $C_6 > 0$ such that

$$\begin{aligned} \|\tilde{Y}_{\phi_i}\|_{\max} &\leq \|Y_i\|_{\max} + \|Y_i - \tilde{Y}_{\phi_i}\|_{\max} \\ &\leq C_6 + C_5 h^4 =: C_7 \end{aligned} \quad (20)$$

is bounded for all for $i = 1, \dots, N$, where $Nh = H$, and we have used that $h \leq 1$.

Now we proceed with the second task and first show that

$$\|\tilde{Y}_{\phi_{i+1}} - \tilde{Y}_{i+1}\|_{\max} \leq 2\|\tilde{Y}_{\phi_i} - \tilde{Y}_i\|_{\max} + C_8 h^5. \quad (21)$$

Comparing the formulas (15) and (17), we have, using $\|AB\|_{\max} \leq n\|A\|_{\max}\|B\|_{\max}$,

$$\begin{aligned} \|k_1 - k'_1\|_{\max} &= h \left\| Df(\phi_i)\tilde{Y}_{\phi_i} - Df(\tilde{\phi}_i)\tilde{Y}_i \right\|_{\max} \\ &\leq hn \left\| Df(\phi_i) - Df(\tilde{\phi}_i) \right\|_{\max} \left\| \tilde{Y}_{\phi_i} \right\|_{\max} \\ &\quad + hn \left\| Df(\tilde{\phi}_i) \right\|_{\max} \left\| \tilde{Y}_{\phi_i} - \tilde{Y}_i \right\|_{\max}. \end{aligned}$$

Using that Df is locally Lipschitz continuous and thus has a Lipschitz constant L_{Df} on the compact convex

hull of S and that the solution ϕ and its numerical approximation $\tilde{\phi}$ starting at $\xi \in S$ lie in the convex hull, we obtain $\left\| D\mathbf{f}(\phi_i) - D\mathbf{f}(\tilde{\phi}_i) \right\|_{\max} \leq L_{Df}C_4h^4$ by (13). We define $L_\nu = \max_{x \in S} \|D\mathbf{f}(x)\|_{\max}$. Altogether, we have

$$\begin{aligned} \|k_1 - k'_1\|_{\max} &\leq L_{Df}nC_4h^5 \left\| \tilde{Y}_{\phi_i} \right\|_{\max} \\ &\quad + hnL_\nu \left\| \tilde{Y}_{\phi_i} - \tilde{Y}_i \right\|_{\max}. \end{aligned}$$

Now we proceed in a similar way with k_2 . From the formula (15) we have $\|k_1\|_{\max} \leq nhL_\nu \|\tilde{Y}_{\phi_i}\|_{\max}$ and thus

$$\begin{aligned} \|k_2 - k'_2\|_{\max} &= \\ h \left\| D\mathbf{f}(\phi_{i+\frac{1}{2}}) \left(\tilde{Y}_{\phi_i} + \frac{k_1}{2} \right) - D\mathbf{f}(\tilde{\phi}_{i+\frac{1}{2}}) \left(\tilde{Y}_i + \frac{k'_1}{2} \right) \right\|_{\max} \\ &\leq h \left\| D\mathbf{f}(\phi_{i+\frac{1}{2}}) \tilde{Y}_{\phi_i} - D\mathbf{f}(\tilde{\phi}_{i+\frac{1}{2}}) \tilde{Y}_i \right\|_{\max} \\ &\quad + \frac{1}{2} h \left\| D\mathbf{f}(\phi_{i+\frac{1}{2}}) k_1 - D\mathbf{f}(\tilde{\phi}_{i+\frac{1}{2}}) k'_1 \right\|_{\max} \\ &\leq hn \left\| D\mathbf{f}(\phi_{i+\frac{1}{2}}) - D\mathbf{f}(\tilde{\phi}_{i+\frac{1}{2}}) \right\|_{\max} \left\| \tilde{Y}_{\phi_i} \right\|_{\max} \\ &\quad + hn \left\| D\mathbf{f}(\tilde{\phi}_{i+\frac{1}{2}}) \right\|_{\max} \left\| \tilde{Y}_{\phi_i} - \tilde{Y}_i \right\|_{\max} \\ &\quad + \frac{1}{2} hn \left\| D\mathbf{f}(\phi_{i+\frac{1}{2}}) - D\mathbf{f}(\tilde{\phi}_{i+\frac{1}{2}}) \right\|_{\max} \|k_1\|_{\max} \\ &\quad + \frac{1}{2} hn \left\| D\mathbf{f}(\tilde{\phi}_{i+\frac{1}{2}}) \right\|_{\max} \|k_1 - k'_1\|_{\max} \\ &\leq L_{Df}nC_4h^5 \left\| \tilde{Y}_{\phi_i} \right\|_{\max} + hnL_\nu \left\| \tilde{Y}_{\phi_i} - \tilde{Y}_i \right\|_{\max} \\ &\quad + \frac{1}{2} L_{Df}n^2C_4h^6L_\nu \left\| \tilde{Y}_{\phi_i} \right\|_{\max} \\ &\quad + \frac{1}{2} hnL_\nu \left[L_{Df}nC_4h^5 \left\| \tilde{Y}_{\phi_i} \right\|_{\max} + hnL_\nu \left\| \tilde{Y}_{\phi_i} - \tilde{Y}_i \right\|_{\max} \right] \\ &\leq C_9h^5 \left\| \tilde{Y}_{\phi_i} \right\|_{\max} + C_{10}h \left\| \tilde{Y}_{\phi_i} - \tilde{Y}_i \right\|_{\max}, \end{aligned}$$

where we have used $h \leq 1$ and new constants.

Now we proceed in a similar way with k_3 ; note that by the formula (15) we have

$$\|k_2\|_{\max} \leq nhL_\nu \left(1 + \frac{L_\nu hn}{2} \right) \left\| \tilde{Y}_{\phi_i} \right\|_{\max}$$

and thus

$$\begin{aligned} \|k_3 - k'_3\|_{\max} &= \\ h \left\| D\mathbf{f}(\phi_{i+\frac{1}{2}}) \left(\tilde{Y}_{\phi_i} + \frac{k_2}{2} \right) - D\mathbf{f}(\tilde{\phi}_{i+\frac{1}{2}}) \left(\tilde{Y}_i + \frac{k'_2}{2} \right) \right\|_{\max} \\ &\leq h \left\| D\mathbf{f}(\phi_{i+\frac{1}{2}}) \tilde{Y}_{\phi_i} - D\mathbf{f}(\tilde{\phi}_{i+\frac{1}{2}}) \tilde{Y}_i \right\|_{\max} \\ &\quad + \frac{1}{2} h \left\| D\mathbf{f}(\phi_{i+\frac{1}{2}}) k_2 - D\mathbf{f}(\tilde{\phi}_{i+\frac{1}{2}}) k'_2 \right\|_{\max} \end{aligned}$$

$$\begin{aligned} &\leq hn \left\| D\mathbf{f}(\phi_{i+\frac{1}{2}}) - D\mathbf{f}(\tilde{\phi}_{i+\frac{1}{2}}) \right\|_{\max} \left\| \tilde{Y}_{\phi_i} \right\|_{\max} \\ &\quad + hn \left\| D\mathbf{f}(\tilde{\phi}_{i+\frac{1}{2}}) \right\|_{\max} \left\| \tilde{Y}_{\phi_i} - \tilde{Y}_i \right\|_{\max} \\ &\quad + \frac{1}{2} hn \left\| D\mathbf{f}(\phi_{i+\frac{1}{2}}) - D\mathbf{f}(\tilde{\phi}_{i+\frac{1}{2}}) \right\|_{\max} \|k_2\|_{\max} \\ &\quad + \frac{1}{2} hn \left\| D\mathbf{f}(\tilde{\phi}_{i+\frac{1}{2}}) \right\|_{\max} \|k_2 - k'_2\|_{\max} \\ &\leq L_{Df}nC_4h^5 \left\| \tilde{Y}_{\phi_i} \right\|_{\max} + hnL_\nu \left\| \tilde{Y}_{\phi_i} - \tilde{Y}_i \right\|_{\max} \\ &\quad + \frac{1}{2} L_{Df}n^2C_4h^6L_\nu \left(1 + \frac{L_\nu hn}{2} \right) \left\| \tilde{Y}_{\phi_i} \right\|_{\max} \\ &\quad + \frac{1}{2} hnL_\nu \left[C_9h^5 \left\| \tilde{Y}_{\phi_i} \right\|_{\max} + C_{10}h \left\| \tilde{Y}_{\phi_i} - \tilde{Y}_i \right\|_{\max} \right] \\ &\leq C_{11}h^5 \left\| \tilde{Y}_{\phi_i} \right\|_{\max} + C_{12}h \left\| \tilde{Y}_{\phi_i} - \tilde{Y}_i \right\|_{\max}, \end{aligned}$$

where we have used $h \leq 1$ and new constants.

Finally, for k_4 , where by (15) we have

$$\|k_3\|_{\max} \leq nhL_\nu \left(1 + \frac{L_\nu hn}{2} + \frac{L_\nu^2 h^2 n^2}{4} \right) \left\| \tilde{Y}_{\phi_i} \right\|_{\max}$$

and thus

$$\begin{aligned} \|k_4 - k'_4\|_{\max} &= \\ h \left\| D\mathbf{f}(\phi_{i+1}) \left(\tilde{Y}_{\phi_i} + k_3 \right) - D\mathbf{f}(\tilde{\phi}_{i+1}) \left(\tilde{Y}_i + k'_3 \right) \right\|_{\max} \\ &\leq h \left\| D\mathbf{f}(\phi_{i+1}) \tilde{Y}_{\phi_i} - D\mathbf{f}(\tilde{\phi}_{i+1}) \tilde{Y}_i \right\|_{\max} \\ &\quad + h \left\| D\mathbf{f}(\phi_{i+1}) k_3 - D\mathbf{f}(\tilde{\phi}_{i+1}) k'_3 \right\|_{\max} \\ &\leq hn \left\| D\mathbf{f}(\phi_{i+1}) - D\mathbf{f}(\tilde{\phi}_{i+1}) \right\|_{\max} \left\| \tilde{Y}_{\phi_i} \right\|_{\max} \\ &\quad + hn \left\| D\mathbf{f}(\tilde{\phi}_{i+1}) \right\|_{\max} \left\| \tilde{Y}_{\phi_i} - \tilde{Y}_i \right\|_{\max} \\ &\quad + hn \left\| D\mathbf{f}(\phi_{i+1}) - D\mathbf{f}(\tilde{\phi}_{i+1}) \right\|_{\max} \|k_3\|_{\max} \\ &\quad + hn \left\| D\mathbf{f}(\tilde{\phi}_{i+1}) \right\|_{\max} \|k_3 - k'_3\|_{\max} \\ &\leq L_{Df}nC_4h^5 \left\| \tilde{Y}_{\phi_i} \right\|_{\max} + hnL_\nu \left\| \tilde{Y}_{\phi_i} - \tilde{Y}_i \right\|_{\max} \\ &\quad + L_{Df}n^2C_4h^6L_\nu \left(1 + \frac{L_\nu hn}{2} + \left(\frac{L_\nu hn}{2} \right)^2 \right) \left\| \tilde{Y}_{\phi_i} \right\|_{\max} \\ &\quad + hnL_\nu \left(C_{11}h^5 \left\| \tilde{Y}_{\phi_i} \right\|_{\max} + C_{12}h \left\| \tilde{Y}_{\phi_i} - \tilde{Y}_i \right\|_{\max} \right) \\ &\leq C_{13}h^5 \left\| \tilde{Y}_{\phi_i} \right\|_{\max} + C_{14}h \left\| \tilde{Y}_{\phi_i} - \tilde{Y}_i \right\|_{\max}, \end{aligned}$$

where we have used $h \leq 1$ and new constants.

Putting these parts together, one can see that

$$\begin{aligned} \left\| \tilde{Y}_{\phi_{i+1}} - \tilde{Y}_{i+1} \right\|_{\max} &\leq \left\| \tilde{Y}_{\phi_i} - \tilde{Y}_i \right\|_{\max} \\ &\quad + \frac{1}{6} \|k_1 - k'_1 + 2(k_2 - k'_2) + 2(k_3 - k'_3) + k_4 - k'_4\|_{\max} \end{aligned}$$

$$\begin{aligned} &\leq \left\| \tilde{Y}_{\Phi_i} - \tilde{Y}_i \right\|_{\max} \left[1 + (nL_v + 2C_{10} + 2C_{12} + C_{14}) \frac{h}{6} \right] \\ &+ \left(L_{D_f} n C_4 + 2C_9 + 2C_{11} + C_{13} \right) \frac{h^5}{6} \left\| \tilde{Y}_{\Phi_i} \right\|_{\max} \\ &\leq 2 \left\| \tilde{Y}_{\Phi_i} - \tilde{Y}_i \right\|_{\max} + C_8 h^5, \end{aligned}$$

where we define

$$C_8 := \frac{C_7}{6} \left(L_{D_f} n C_4 + 2C_9 + 2C_{11} + C_{13} \right),$$

and we have used (20) and the fact that

$$h \leq \frac{6}{nL_v + 2C_{10} + 2C_{12} + C_{14}} =: h^*. \quad (22)$$

This shows (21), which in turn shows with

$$\left\| \tilde{Y}_{\Phi_i} - \tilde{Y}_i \right\|_{\max} = 0 \quad \text{for } i = 0$$

and by iteration that

$$\begin{aligned} \left\| \tilde{Y}_{\Phi_i} - \tilde{Y}_i \right\|_{\max} &\leq (2^i - 1) C_8 h^5 \\ &\leq 7C_8 h^5 =: C_{15} h^5. \end{aligned} \quad (23)$$

for $i = 1, 2, 3$.

Now the initializing steps with RK4 are finished. For the AB4 method we follow a similar idea. For $i \geq 3$ we have

$$\begin{aligned} &\left\| \tilde{Y}_{\Phi_{i+1}} - \tilde{Y}_{i+1} \right\|_{\max} \leq \left\| \tilde{Y}_{\Phi_i} - \tilde{Y}_i \right\|_{\max} \\ &+ \frac{55h}{24} \left\| Df(\Phi_i) \tilde{Y}_{\Phi_i} - Df(\tilde{\Phi}_i) \tilde{Y}_i \right\|_{\max} \\ &+ \frac{59h}{24} \left\| Df(\Phi_{i-1}) \tilde{Y}_{\Phi_{i-1}} - Df(\tilde{\Phi}_{i-1}) \tilde{Y}_{i-1} \right\|_{\max} \\ &+ \frac{37h}{24} \left\| Df(\Phi_{i-2}) \tilde{Y}_{\Phi_{i-2}} - Df(\tilde{\Phi}_{i-2}) \tilde{Y}_{i-2} \right\|_{\max} \\ &+ \frac{9h}{24} \left\| Df(\Phi_{i-3}) \tilde{Y}_{\Phi_{i-3}} - Df(\tilde{\Phi}_{i-3}) \tilde{Y}_{i-3} \right\|_{\max} \end{aligned}$$

Using the estimate

$$\begin{aligned} &\left\| Df(\Phi_i) \tilde{Y}_{\Phi_i} - Df(\tilde{\Phi}_i) \tilde{Y}_i \right\|_{\max} \\ &\leq n \left\| Df(\Phi_i) - Df(\tilde{\Phi}_i) \right\|_{\max} \left\| \tilde{Y}_{\Phi_i} \right\|_{\max} \\ &+ n \left\| Df(\tilde{\Phi}_i) \right\|_{\max} \left\| \tilde{Y}_{\Phi_i} - \tilde{Y}_i \right\|_{\max} \\ &\leq nL_{D_f} C_4 C_7 h^4 + nL_v \left\| \tilde{Y}_{\Phi_i} - \tilde{Y}_i \right\|_{\max}, \end{aligned}$$

similarly to the argumentation above for RK4, for each of the terms $i, i-1, i-2, i-3$ we obtain

$$\begin{aligned} &\left\| \tilde{Y}_{\Phi_{i+1}} - \tilde{Y}_{i+1} \right\|_{\max} \leq \left\| \tilde{Y}_{\Phi_i} - \tilde{Y}_i \right\|_{\max} \quad (24) \\ &+ \frac{20}{3} nL_{D_f} C_4 C_7 h^5 + hnL_v \left(\frac{55}{24} \left\| \tilde{Y}_{\Phi_i} - \tilde{Y}_i \right\|_{\max} \right. \\ &+ \frac{59}{24} \left\| \tilde{Y}_{\Phi_{i-1}} - \tilde{Y}_{i-1} \right\|_{\max} \\ &\left. + \frac{37}{24} \left\| \tilde{Y}_{\Phi_{i-2}} - \tilde{Y}_{i-2} \right\|_{\max} + \frac{9}{24} \left\| \tilde{Y}_{\Phi_{i-3}} - \tilde{Y}_{i-3} \right\|_{\max} \right). \end{aligned}$$

We have assumed that

$$h \leq h^{**} := \frac{3}{20nL_v} \quad (25)$$

and want to show that

$$\begin{aligned} \left\| \tilde{Y}_{\Phi_i} - \tilde{Y}_i \right\|_{\max} &\leq C_{16} 2^N h^5 = C_{16} 2^H h^4 \frac{h}{2^h} \\ &\leq C_{16} 2^H h^4 \end{aligned} \quad (26)$$

for $i = 1, \dots, N$; here $C_{16} := \max(C_{15}, \frac{20}{3} nL_{D_f} C_4 C_7)$ and we have used that $\frac{h}{2^h} \leq 1$ for all $h \geq 0$. Hence, this shows that we have a global estimate of order 4 with a constant depending on H .

To show (26) we denote $a_i = \left\| \tilde{Y}_{\Phi_i} - \tilde{Y}_i \right\|_{\max}$ and prove that

$$a_i \leq b_i := \hat{C}(2^i - 1),$$

where $\hat{C} = C_{16} h^5$, for all $i = 0, \dots, N$. Note that b_i is the solution of the iteration $b_0 = 0$ and

$$b_{i+1} = 2b_i + \hat{C}.$$

We will now show $a_i \leq b_i$ by induction with respect to i . For $i = 0, 1, 2, 3$ this follows from (23). Now we assume that for $i \geq 3$ the inequality $a_j \leq b_j$ holds for all $j = 0, \dots, i$ and we show it for $i+1$: by (24) we have

$$\begin{aligned} a_{i+1} &\leq a_i + \frac{20}{3} nL_{D_f} C_4 C_7 h^5 \\ &+ hnL_v \left(\frac{55}{24} a_i + \frac{59}{24} a_{i-1} + \frac{37}{24} a_{i-2} + \frac{9}{24} a_{i-3} \right) \\ &\leq b_i + \frac{20}{3} nL_{D_f} C_4 C_7 h^5 + hnL_v \frac{20}{3} b_i \\ &\leq 2b_i + \hat{C} = b_{i+1}, \end{aligned}$$

where we have used $h \leq 3/(20nL_v)$, $h \leq 1$ and the induction assumption. This shows the induction and thus (26).

Finally, we obtain for the error, using (19) and (26), that

$$\begin{aligned} \left\| Y_i - \tilde{Y}_i \right\|_{\max} &\leq \left\| Y_i - \tilde{Y}_{\Phi_i} \right\|_{\max} + \left\| \tilde{Y}_{\Phi_i} - \tilde{Y}_i \right\|_{\max} \\ &\leq C_5 h^4 + C_{16} 2^H h^4. \end{aligned} \quad (27)$$

for all $i = 1, \dots, N$, hence a global error of order 4.

The algorithm to approximate the solutions to the initial-value problems (3) and (4) can now be summarized as:

1. Fix the time-horizon H and the number of time-steps N .
2. For the initiation phase for the AB4 multi-step method, fix $\tilde{x}_0 = \xi$ and set $h_0 = \frac{1}{2} H/N$.
3. Use the RK4 formula (9) with $h = h_0$ to compute \tilde{x}_{i+1} for $i = 0, 1, 2, 3, 4, 5$.

4. Relabel the solution terms using $\tilde{\phi}_{i/2} = \tilde{\mathbf{x}}_i$ for $i = 0, 1, \dots, 6$, e.g. $\tilde{\phi}_{\frac{1}{2}} = \tilde{\mathbf{x}}_1$, $\tilde{\phi}_1 = \tilde{\mathbf{x}}_2$ etc.
5. Set $\tilde{Y}_0 = I$, $h = H/N$, and use the RK4 formula (15) to compute \tilde{Y}_{i+1} for $i = 0, 1, 2$.
6. Now the initialization phase for the AB4 method is over and we have $\tilde{\phi}_i$ and \tilde{Y}_i at our disposal for $i = 0, 1, 2, 3$. Set $h = H/N$ and use formulas (10) and (18) for $i = 3, 4, \dots, N-1$ to compute the remaining $\tilde{\phi}_i$ and \tilde{Y}_i .

Remark 3.1. *The two following observations are useful for our approach.*

- a. *Note that since we perform the computations on a compact set S , that is positively invariant for both the system (1) and the numerical integrator, we have finite upper bounds on the absolute values of continuous s -th order derivatives of the components of \mathbf{f} if \mathbf{f} is C^s on the convex hull of S . Thus, for $s = 5$ we can choose C_j for $j = 1, 2, \dots, 5$ and also L_v as uniform constants for the whole interval $[0, H]$ independent of the time step t_i and $\xi \in S$.*
- b. *It is not necessary to keep track of more than just the four most recent values of $\tilde{\phi}_i$ and \tilde{Y}_i in step 6 to use formulas (10) and (18).*

3.3 Step III: Numerical Quadrature

The numerical quadrature of formula (2) can be done on the fly as explained below. Effectively, this can be implemented by interpreting i modulo 4. Since the numerical solutions to (3) and (4) are $O(h^4)$, there is little additional accuracy gained by using a higher-order formula than $O(h^4)$ for the numerical quadrature of (2). However, by using the Composite Trapezoidal Rule and a Romberg-like extrapolation, one can use an $O(h^{2(p+1)})$ method with negligible overhead, where $N = 2^p q$, $p, q \in \mathbb{N}_+$, and h is the step size for both the numerical integration and the numerical quadrature. This method and its implementation is explained in detail in (Hafstein, 2019, Sec. III). We review the essential parts here.

Approximations to the solutions of system (3) and system (4) with a particular initial value $\phi(0, \xi) = \xi$ and $Y(0) = I$ are computed at $N+1$ equally distributed time points on the time interval $[0, H]$:

$$\tilde{\phi}_i \approx \phi(t_i, \xi) \text{ and } \tilde{Y}_i \approx Y(t_i) \text{ at } t_i = \frac{iH}{N} \quad (28)$$

for $i = 0, 1, \dots, N$, where $N = 2^p q$, $p, q \in \mathbb{N}_+$; how this is done and how accurate these values are was explained in Step II.

We first consider the quadrature rule error assuming that we use the true values for ϕ and Y ; later we

consider the error caused by using the approximate values $\tilde{\phi}$ and \tilde{Y} . We split the integral in (2) into two parts

$$M(\xi) = I_\xi + \int_H^\infty Y(\tau)^T C(\phi(\tau, \mathbf{x})) Y(\tau) d\tau, \text{ where}$$

$$I_\xi := \int_0^H Y(\tau)^T C(\phi(\tau, \mathbf{x})) Y(\tau) d\tau,$$

and I_ξ is approximated by $\tilde{M}(\xi)$ using formula (5) and a variant of the Romberg integration. Set

$$\alpha(\tau, \xi) = Y(\tau)^T C(\phi(\tau, \xi)) Y(\tau) \quad (29)$$

and $\alpha_i := \alpha(t_i, \xi)$ for $i = 0, 1, \dots, N$. First, we use the composite Trapezoidal rule to approximate $I_\xi := M(\xi)$ using $N, N/2, \dots, q$ intervals. For this, define recursively $N_0 := N$ and $N_{k+1} := N_k/2$ for $k = 0, \dots, p-1$; note that $N_p = q$. Define $h'_k := H/N_k$ for $k = 0, \dots, p$. We set

$$\text{Trap}_k = h'_k \left(\frac{\alpha_0 + \alpha_{N_R}}{2} + \sum_{j=1}^{N_k-1} \alpha_{j2^k} \right) \quad (30)$$

for $k = 0, 1, \dots, p$. It is well known, cf. e.g. (Bauer et al., 1963), that if the integrand is $C^{2(p+1)}$, which is the case if $\mathbf{f} \in C^{2(p+1)+1}$ and $C \in C^{2(p+1)}$, that by extrapolation using the tableau

$$R_{r,0} := \text{Trap}_r \text{ for } r = 0, 1, \dots, p \quad (31)$$

and then for $s = 1, \dots, p$,

$$R_{r,s} = \frac{4^s R_{r,s-1} - R_{r+1,s-1}}{4^s - 1} \text{ for } 0 \leq r \leq p-s, \quad (32)$$

we get that

$$\|R_{0,p} - I_\xi\|_{\max} \leq C_R H h^{2(p+1)} \max_{t \in [0, H]} \left\| \frac{\partial^{2(p+1)}}{\partial t^{2(p+1)}} \alpha(t, \xi) \right\|_{\max}$$

for a constant C_R , independent of H and α . Here $h = h'_0 = H/N = H/(2^p q)$ is the length of the interval between two consecutive time-points the solution is computed at.

Finally, we substitute the values α_i by $\tilde{\alpha}_i := \tilde{Y}_i^T C(\tilde{\phi}_i) \tilde{Y}_i$ into the formulas (30) and denote the corresponding $R_{r,s}$ by $\tilde{R}_{r,s}$. By the estimates (27) and (13) and because C is Lipschitz on S we have for a fixed $H > 0$ that $\|\alpha_i - \tilde{\alpha}_i\|_{\max} \leq C_I h^4$, where $C_I > 0$ is a constant that can be chosen independently of $\xi \in S$. Further, $R_{r,s} = \sum_{i=0}^N \lambda_i \alpha_i$ where $\lambda_i \geq 0$ and $\sum_{i=0}^N \lambda_i = 1$, see (Bauer et al., 1963), and thus $\|R_{r,s} - \tilde{R}_{r,s}\|_{\max} \leq (C_R + C_I) h^4$ independent of $\xi \in S$ if $p \geq 1$. Thus, there is a constant C_H independent of $h > 0$ and $\xi \in S$, but dependent on H , such that

$$\|I_\xi - \tilde{R}_{r,s}\|_{\max} \leq C_H h^4 \quad (33)$$

for $h \leq \min(h', 1, h^*, h^{**})$.

To summarize all these discussions, let us provide the error estimate for the numerical computation of the contraction metric M .

4 MAIN RESULT

After all the preparation in the last two sections, we are ready to prove the main result of this work.

Theorem 4.1 (Error estimate). *Assume \mathbf{f} in (1) is $C^{2(p+1)+1}$ and C is $C^{2(p+1)}$ for an integer $p \geq 1$ and let M be defined by formula (2) on $\mathcal{A}(\mathbf{x}_0)$, i.e.*

$$M(\boldsymbol{\xi}) = \int_0^\infty \boldsymbol{\psi}(\tau, \boldsymbol{\xi})^T C(\boldsymbol{\phi}(\tau, \boldsymbol{\xi})) \boldsymbol{\psi}(\tau, \boldsymbol{\xi}) d\tau.$$

Then, for any compact $K \subset \mathcal{A}(\mathbf{x}_0)$ and $\varepsilon > 0$, there exists $H^* > 0$ such that for all fixed and finite $H \geq H^*$ there exist $N^* = 2^p q^*$, $p, q^* \in \mathbb{N}_+$, such that for all $N = 2^p q$, $q \geq q^*$, we have

$$\|M(\boldsymbol{\xi}) - \tilde{M}(\boldsymbol{\xi})\|_{\max} \leq \varepsilon$$

for all $\boldsymbol{\xi} \in K$. Here $h := H/N$ and $\tilde{M}(\boldsymbol{\xi})$ is the result of the numerical method, i.e. the matrix $\tilde{R}_{0,p}$, computed as described in Section 3 with initial value $\tilde{\boldsymbol{\xi}}$, and using the interval H and the approximations $\tilde{\boldsymbol{\phi}}$ and \tilde{Y} with step-size h in the numerical integration and quadrature.

Proof. Let $K \subset \mathcal{A}(\mathbf{x}_0)$ and $\varepsilon > 0$ be given. By (Giesl et al., 2023d, Thms. 2.1 and 3.5) there exists a compact $S \subset \mathbb{R}^n$, $K \subset S \subset \mathcal{A}(\mathbf{x}_0)$ and $h' > 0$, such that S is positively invariant for system (1) and the AB4 method initialized with RK4 to approximate its trajectories for all step sizes $0 < h \leq h'$, i.e. $\tilde{\boldsymbol{\phi}}_i \in S$ for all $i \in \mathbb{N}_0$ if $\tilde{\boldsymbol{\phi}}_0 = \boldsymbol{\xi} \in S$ and the step-size is h .

The proof of (Giesl and Wendland, 2019, Thm. 2.2, (13), (14), and (15)) implies that there are constants $d, d_1, d_2, \rho > 0$ such that

$$\|\boldsymbol{\psi}(\tau, \boldsymbol{\xi})\|_{\max} \leq d_1 e^{-\rho\tau} \quad (34)$$

$$\|C(\boldsymbol{\phi}(\tau, \boldsymbol{\xi}))\|_{\max} \leq d_2 \quad (35)$$

$$\|\boldsymbol{\psi}(\tau, \boldsymbol{\xi})^T C(\boldsymbol{\phi}(\tau, \boldsymbol{\xi})) \boldsymbol{\psi}(\tau, \boldsymbol{\xi})\|_{\max} \leq d e^{-2\rho\tau} \quad (36)$$

for all $\tau \geq 0$ and all $\boldsymbol{\xi} \in S$. Note that the proof also holds for the case of a compact subset of $\mathcal{A}(\mathbf{x}_0)$. The reason is that a compact subset of the basin of attraction is uniformly attracted to \mathbf{x}_0 , which implies that uniform constants can be chosen in (Giesl and Wendland, 2019, (13), (14), and (15)) which only depend on S . Let $H^* > 0$ be so large that

$$\left\| \int_{H^*}^\infty \boldsymbol{\psi}(\tau, \boldsymbol{\xi})^T C(\boldsymbol{\phi}(\tau, \boldsymbol{\xi})) \boldsymbol{\psi}(\tau, \boldsymbol{\xi}) d\tau \right\|_{\max} \leq \frac{d}{2\rho} e^{-2\rho H^*} \leq \frac{\varepsilon}{2}. \quad (37)$$

Fix some $H \geq H^*$. Choose $N^* = 2^p q^*$ so large that both $\tilde{h} := H/N^* \leq \min(h', 1)$ and $C_H(\tilde{h})^4 \leq \frac{\varepsilon}{2}$ hold, where C_H is the constant from (33). Then we have

$$\left\| \int_0^H \boldsymbol{\psi}(\tau, \boldsymbol{\xi})^T C(\boldsymbol{\phi}(\tau, \boldsymbol{\xi})) \boldsymbol{\psi}(\tau, \boldsymbol{\xi}) d\tau - \tilde{R}_{0,p} \right\|_{\max} \leq C_H(\tilde{h})^4 \leq \frac{\varepsilon}{2} \quad (38)$$

at any point $\boldsymbol{\xi} \in K \subset \mathcal{A}(\mathbf{x}_0) \subset \mathbb{R}^n$. Here $\tilde{R}_{0,p}$ is the Romberg approximation of the integral

$$\int_0^H \alpha(\tau, \boldsymbol{\xi}) d\tau,$$

with α defined in (29), where the values $\alpha_i := \alpha(t_i, \boldsymbol{\xi})$, $t_i := i\tilde{h}$, have been substituted by the approximations $\tilde{\alpha}_i$, computed using the numerical solutions $\tilde{\boldsymbol{\phi}}$ and \tilde{I} to $\tau \mapsto \boldsymbol{\phi}(\tau, \mathbf{x})$ and $I(\tau) = \boldsymbol{\psi}(\tau, \mathbf{x})$ as described in Step I and Step II and using step-size \tilde{h} . For any step-size $h := H/N$, $N = 2^p q$ where $q \geq q^*$ is an integer, an analogous estimate holds with \tilde{h} replaced by $h \leq \tilde{h}$ and the proposition follows. \square

5 CONCLUSIONS

We have shown that contraction metrics for systems $\dot{\mathbf{x}} = \mathbf{f}(\mathbf{x})$ with an exponentially stable equilibrium \mathbf{x}_0 can be computed with arbitrary precision. In particular, we have proven that the error in the computations can be uniformly bounded on compact subsets of the basin of attraction $\mathcal{A}(\mathbf{x}_0)$. The bound is uniform in the sense that given a compact subset $K \subset \mathcal{A}(\mathbf{x}_0)$ and an $\varepsilon > 0$, we can choose the parameters of the numerical method, i.e. the time interval H and the step size h , such that the difference between the computed metric $\tilde{M}(\boldsymbol{\xi})$ and the metric $M(\boldsymbol{\xi})$ from formula (2) is bounded by ε for all $\boldsymbol{\xi} \in K$. This result advances the error analysis of the computation of contraction metrics, and justifies the use of these methods in practice.

ACKNOWLEDGEMENT

The research done for this paper was partially supported by the Icelandic Research Fund in the grant 228725-051, *Linear Programming as an Alternative to LMIs for Stability Analysis of Switched Systems*, which is gratefully acknowledged.

REFERENCES

- Aminzare, Z. and Sontag, E. (2014). Contraction methods for nonlinear systems: A brief introduction and some open problems. In *Proceedings of the 53rd IEEE Conference on Decision and Control*, pages 3835–3847.
- Aylward, E., Parrilo, P., and Slotine, J.-J. (2008). Stability and robustness analysis of nonlinear systems via contraction metrics and SOS programming. *Automatica*, 44(8):2163–2170.
- Bauer, F., Rutishauser, H., and Stiefel, E. (1963). New aspects in numerical quadrature. In *Proceedings of Symposia in Applied Mathematics: Experimental Arithmetic, High Speed Computing and Mathematics*, volume 15, pages 199–218. AMS.
- Borg, G. (1960). *A condition for the existence of orbitally stable solutions of dynamical systems*. Kungl. Tekn. Högsk. Handl. 153.
- Demidovič, B. (1961). On the dissipativity of a certain non-linear system of differential equations. I. *Vestnik Moskov. Univ. Ser. I Mat. Meh.*, 1961(6):19–27.
- Deuffhard, P. and Hohmann, A. (2008). *Numerische Mathematik 2*. de Gruyter, 4th edition.
- Forni, F. and Sepulchre, R. (2014). A differential Lyapunov framework for Contraction Analysis. *IEEE Transactions on Automatic Control*, 59(3):614–628.
- Giesl, P. (2015). Converse theorems on contraction metrics for an equilibrium. *J. Math. Anal. Appl.*, (424):1380–1403.
- Giesl, P. (2019). Computation of a contraction metric for a periodic orbit using meshfree collocation. *SIAM J. Appl. Dyn. Syst.*, 18(3):1536–1564.
- Giesl, P. and Hafstein, S. (2013). Construction of a CPA contraction metric for periodic orbits using semidefinite optimization. *Nonlinear Anal.*, 86:114–134.
- Giesl, P., Hafstein, S., Haraldsdottir, M., Thorsteinsson, D., and Kawan, C. (2023a). Subgradient algorithm for computing contraction metrics for equilibria. *J. Comput. Dynamics*, 10(2):281–303.
- Giesl, P., Hafstein, S., and Kawan, C. (2023b). Review on contraction analysis and computation of contraction metrics. *J. Comput. Dynamics*, 10(1):1–47.
- Giesl, P., Hafstein, S., and Mehrabinezhad, I. (2021a). Computation and verification of contraction metrics for exponentially stable equilibria. *J. Comput. Appl. Math.*, 390:Paper No. 113332.
- Giesl, P., Hafstein, S., and Mehrabinezhad, I. (2021b). Computation and verification of contraction metrics for periodic orbits. *J. Math. Anal. Appl.*, 503(2):Paper No. 125309, 32.
- Giesl, P., Hafstein, S., and Mehrabinezhad, I. (2023c). Contraction metric computation using numerical integration and quadrature. *Submitted*.
- Giesl, P., Hafstein, S., and Mehrabinezhad, I. (2023d). Positively invariant sets for ODEs and numerical integration. In *Proceedings of the 20th International Conference on Informatics in Control, Automation and Robotics (ICINCO)*, page (submitted).
- Giesl, P. and Wendland, H. (2019). Construction of a contraction metric by meshless collocation. *Discrete Contin. Dyn. Syst. Ser. B*, 24(8):3843–3863.
- Hafstein, S. (2019). Numerical ODE solvers and integration methods in the computation of CPA Lyapunov functions. In *18th European Control Conference (ECC)*, pages 1136–1141. IEEE.
- Hartman, P. (1961). On stability in the large for systems of ordinary differential equations. *Canadian J. Math.*, 13:480–492.
- Hartman, P. (1964). *Ordinary Differential Equations*. Wiley, New York.
- Krasovskii, N. N. (1963). *Problems of the Theory of Stability of Motion*. Mir, Moskow, 1959. English translation by Stanford University Press.
- Lewis, D. (1951). Differential equations referred to a variable metric. *Amer. J. Math.*, 73:48–58.
- Lewis, D. C. (1949). Metric properties of differential equations. *Amer. J. Math.*, 71:294–312.
- Lohmiller, W. and Slotine, J.-J. (1998). On Contraction Analysis for Non-linear Systems. *Automatica*, 34:683–696.
- Sauer, T. (2012). *Numerical Analysis*. Pearson, 2nd edition.
- Simpson-Porco, J. and Bullo, F. (2014). Contraction theory on Riemannian manifolds. *Systems Control Lett.*, 65:74–80.
- Walter, W. (1998). *Ordinary Differential Equation*. Springer.