Positively invariant sets for ODEs and predictor-corrector multi-step numerical solvers*

Peter Giesl[1 [0000−0003−1421−6980]], Sigurdur Hafstein[2 [0000−0003−0073−2765]], and Iman Mehrabinezhad[3 [0000−0002−6346−9901]]

[1] Department of Mathematics, University of Sussex, Falmer, BN1 9QH, United Kingdom `P.A.Giesl@sussex.ac.uk`

[2] University of Iceland, Faculty of Physical Sciences, Dunhagi 5, 107 Reykjavik, Iceland `shafstein@hi.is`

[3] CARE-C Center, The Cyprus Institute, 20 Konstantinou Kavafi St., 2121, Nicosia, Cyprus `i.mehrabinzhad@cyi.ac.cy`

Abstract. We investigate positively invariant sets for an ordinary differential equation (ODE), that are also positively invariant for numerical methods to compute its solution. In particular, we show that for an ODE with an exponentially stable equilibrium and an arbitrary compact subset of its basin of attraction, we can establish the existence of a larger compact set that is positively invariant for both the ODE, one-step explicit- and multi-step numerical methods, and even predictor-corrector multi-step methods. We demonstrate in an example the use of this method when computing a contraction metric for an ODE with an exponentially stable equilibrium.

Keywords: Positively invariant sets, Ordinary Differential Equations, Numerical Integration, Multi-Step Methods, Predictor-Corrector Integration

1 Introduction

We are concerned with the autonomous ordinary differential equation (ODE)

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

Let $\mathbf{x}_0 \in \mathbb{R}^n$ be an exponentially stable equilibrium of (1) and denote by

$$\mathcal{A}(\mathbf{x}_0) := \{\mathbf{x} \in \mathbb{R}^n \colon \lim_{t\to\infty} \boldsymbol{\phi}(t,\mathbf{x}) = \mathbf{x}_0\}$$

its basin of attraction, where $\boldsymbol{\phi}(t,\boldsymbol{\xi})$ denotes the solution $\mathbf{x}(t)$ to the initial value problem (1) with $\mathbf{x}(0) = \boldsymbol{\xi}$. Note that for fixed $\boldsymbol{\xi}$, the solution $\boldsymbol{\phi}(t,\boldsymbol{\xi})$ is defined in an open interval $t \in (-c_1, c_2)$ with $c_1, c_2 > 0$.

A positively invariant set $S \subset \mathbb{R}^n$ for the ODE (1) is a set such that $\boldsymbol{\phi}([0,\infty), S) \subset S$, i.e. a solution starting in S stays in S for all future times;

* This work was supported in part by the Icelandic Research Fund under Grant 228725-051.

in particular, it is defined for all future times. We will show that for any compact set $K \subset \mathcal{A}(\mathbf{x}_0)$ we can establish the existence of a compact $K \subset S \subset \mathcal{A}(\mathbf{x}_0)$ that is not only positively invariant for (1), but also for numerical methods to approximate its solutions. The latter means, that a numerical method that approximates the solution trajectory $t \mapsto \boldsymbol{\phi}(t, \boldsymbol{\xi})$ through a sequence $\widetilde{\boldsymbol{\phi}}_i \approx \boldsymbol{\phi}(hi, \boldsymbol{\xi})$, where $h > 0$ is a small constant and $i \in \mathbb{N}$, fulfills $\widetilde{\boldsymbol{\phi}}_i \in S$ for all $i \in \mathbb{N}$ if $\widetilde{\boldsymbol{\phi}}_0 = \boldsymbol{\xi} \in S$.

These results are important for various methods to compute Lyapunov functions and contraction metrics for ODEs and, indeed, often necessary to show that these methods always work. In the qualitative analysis of ODEs, Lyapunov functions play a central role and are studied in virtually all textbooks and monographs on ODEs, cf. e.g. [38,44,54,60,62,63]. Recently, there has also been much interest in so-called contraction metrics, which are Riemannian metrics that correspond to Lyapunov functions on the tangent space [23]. Some references to contraction metrics are [1,7,9,13,15,40,41,45,46,47,49,57] and the textbook [8]. The analytical computation of a contraction metric for an ODE, a matrix-valued function, is even more difficult than the computation of a Lyapunov function. For general ODEs it is very difficult to compute a Lyapunov function or a contraction metric analytically and therefore one resorts to numerical methods. Numerical methods for the construction of contraction metrics include [3,16,17,22], see also the recent review [23]. Numerical methods for the computation of Lyapunov functions include the following: the computation of rational Lyapunov functions was studied in [58,59], sum-of-squared (SOS) polynomial Lyapunov functions were computed using semi-definite optimization (SOS method) in [2,51,52], see also [43,53] for other approaches using polynomials, and a Zubov-type PDE was numerically solved using radial basis functions (RBF method) in [14]. For an overview of more methods see the review [20]. Linear programming was used to parameterize continuous and piecewise affine (CPA) Lyapunov functions in [42,50] in the so-called CPA method. In the CPA method the domain of interest, where the Lyapunov function is to be computed, is subdivided into simplices and a system specific feasibility problem is constructed, a feasible solution of which parameterizes a CPA Lyapunov function for the system. In [18,29,30] it was shown that the CPA method always succeeds in computing a Lyapunov function for an ODE with an asymptotically stable equilibrium, provided that the simplices in the triangulation are sufficiently small.

The CPA and the RBF methods were combined in [19] to deliver a method that inherits the numerical efficiency of the RBF method and the rigour of the CPA method. This is accomplished by solving a system of linear equations in the RBF method rather than the linear optimization problem from the CPA method. The thus obtained function is subsequently verified to be a true Lyapunov function by investigating whether it fulfills the constraints of the optimization problem. Using this combined method, one is always able to compute a true Lyapunov function in any compact subset of an exponentially stable equilibrium's basin of attraction.

A similar approach uses numerical integration of solution trajectories of the ODE to generate values for the variables of the feasibility problem for the CPA method and then verifies the constraints, see [4,5,6,11,12,33,34,35,36,37,48] and also [28,31] for more implementation oriented papers. This technique works well in practice and in [21] it is proved that it always works, assuming that one can find a compact set that is positively invariant for not only the ODE in question, but also for the numerical method used to approximate its solution trajectories.

The main contribution of this paper is to assert the existence of such a positively invariant set in Theorems 2 and 4 and to show how it can be computed in Theorem 3. Further, we show that one can use a range of practical numerical methods to approximate the solution trajectories, namely not only one-step explicit methods, but also explicit multi-step methods and even predictor-corrector multi-step methods. The results in this paper are also essential for numerical methods for contraction metrics using numerical integration and quadrature. In [25], the results of this paper are used to derive a uniform error estimate on compact sets, needed to prove in [27] that such an integration-quadrature method always succeeds in computing a contraction metric for an ODE with an exponentially stable equilibrium.

Let us give an overview of the paper: In Section 2 we recall some facts about numerical integration methods of ODEs and prove that a wide range of methods, including multi-step methods, satisfy certain approximation properties, see Definition 1. In Section 3 we establish the existence of positively invariant sets, both for the dynamics of the system (1) and a numerical integration scheme to approximate the solution trajectories in the basin of attraction of an exponentially stable equilibrium; the main result is Theorem 4. Such positively invariant sets are very useful, in fact necessary, to prove that Lyapunov functions and contraction metrics can be approximated arbitrarily close on compact subsets of basins of attraction, using numerical integration with subsequent numerical quadrature. We prove our results using the fourth-order Adams-Bashforth (AB4) multi-step scheme initialized with fourth-order Runge-Kutta (RK4) as well as a predictor-corrector method based on the implicit fourth-order Adams-Moulton (AM4) method, but we discuss how the results can be extended to AB-RK and AM numerical schemes of arbitrary order. Finally, we give an example of our results in Section 4 before we conclude the paper in Section 5.

This paper extends the results presented at the 20$^{\text{th}}$ International Conference on Informatics in Control, Automation and Robotics (ICINCO 2023) and published in the conference proceedings [26], in several different ways. For example, we have extended the previous results to cover general p-th order multistep methods as well as predictor-corrector methods, and we present a new example.

Notation: We define $\mathbb{N}_0 := \{0, 1, 2, \ldots\}$ and $\mathbb{N} := \mathbb{N}_0 \setminus \{0\}$. We denote the usual Euclidian norm on $\mathbb{R}^n$ by $\|\cdot\|_2$. The closure of a set $U \subset \mathbb{R}^n$ is denoted by $\overline{U}$ and its boundary by ∂U. The distance between a point $\mathbf{x} \in \mathbb{R}^n$ and a set $K \subset \mathbb{R}^n$ is represented by $d(\mathbf{x}, K) := \inf_{\mathbf{y} \in K} \|\mathbf{x} - \mathbf{y}\|_2$. Throughout the paper, $\|\cdot\|$ is a fixed, but arbitrary norm on $\mathbb{R}^n$.

2 Numerical Integration Methods

As discussed above, when constructing Lyapunov functions or contraction metrics using numerical integration, one needs to solve many initial value problems for the ODE (1), with initial data located at the vertices of a triangulation of a compact set in $\mathbb{R}^n$. Because the number of vertices can be very large, it is advantageous to use multi-step methods rather than single-step methods, as these are considerably faster for the same degree of precision. We will prove our results for

1. One-step explicit methods, in particular Runge-Kutta (RK) methods.
2. Multi-step explicit methods, in particular Adams-Bashforth (AB) methods.
3. Multi-step predictor-corrector methods, where the predictor step is done using AB and the corrector step using the Adams-Moulton (AM) method.

Let us go through these kind of methods and give examples; for a more detailed discussion see, e.g. [10,39,55] and the references therein.

2.1 One-Step Methods

In a one-step method, $\widetilde{\boldsymbol{\phi}}_{i+1}$ is computed directly from $\widetilde{\boldsymbol{\phi}}_i$. Typical examples are the RK methods. Let us write down the explicit formulas for RK of fourth order (RK4): Fix the *step-size* $h > 0$ and set $\widetilde{\boldsymbol{\phi}}_0 = \boldsymbol{\xi}$. Then, for every $i \in \mathbb{N}_0$ set

$$\begin{aligned} \mathbf{k}_1 &= h\mathbf{f}(\widetilde{\boldsymbol{\phi}}_i) \\ \mathbf{k}_2 &= h\mathbf{f}(\widetilde{\boldsymbol{\phi}}_i + \mathbf{k}_1/2) \\ \mathbf{k}_3 &= h\mathbf{f}(\widetilde{\boldsymbol{\phi}}_i + \mathbf{k}_2/2) \\ \mathbf{k}_4 &= h\mathbf{f}(\widetilde{\boldsymbol{\phi}}_i + \mathbf{k}_3) \\ \widetilde{\boldsymbol{\phi}}_{i+1} &= \widetilde{\boldsymbol{\phi}}_i + \frac{1}{6}(\mathbf{k}_1 + 2\mathbf{k}_2 + 2\mathbf{k}_3 + \mathbf{k}_4). \end{aligned} \tag{2}$$

RK4 is said to be of fourth-order, because it can be shown that if $\mathbf{f}$ in (1) is C^4, then for every compact set $K \subset \mathbb{R}^n$ one can find constants $h^* > 0$ and C_{RK4} such that

$$\|\widetilde{\boldsymbol{\phi}}_1(\boldsymbol{\xi}) - \boldsymbol{\phi}(h, \boldsymbol{\xi})\| \le C_{\text{RK4}} h^5 \tag{3}$$

for all step-sizes $0 < h \le h^*$. The constant C_{RK4} depends on the derivatives of $\mathbf{f}$, up to and including the fourth order, in a compact and convex set $\widetilde{K} \supset K$. The set $\widetilde{K}$ is not really of interest and a large enough set clearly exists; it only has to be large enough to contain all points in the Taylor polynomial expansions of $\mathbf{f}$ and $\boldsymbol{\phi}$ needed to make the appropriate estimates, e.g.

$$\boldsymbol{\phi}(h, \boldsymbol{\xi}), \widetilde{\boldsymbol{\phi}}_1(\boldsymbol{\xi}), \boldsymbol{\xi} + \mathbf{k}_1/2, \boldsymbol{\xi} + \mathbf{k}_2/2, \boldsymbol{\xi} + \mathbf{k}_3 \quad \text{for all} \quad \boldsymbol{\xi} \in K.$$

The RK4 method is said to be of fourth order, although we have $\mathcal{O}(h^5)$ in (3), the so-called *local truncation error*, because one power of h is lost when one goes from one-step to many steps.

Remark 1. It is necessary to fix an upper bound $h^* > 0$ for the general case, because solutions $t \mapsto \boldsymbol{\phi}(t, \boldsymbol{\xi})$ might blow-up in a finite time. As a simple example consider the one-dimensional ODE $\dot{x} = f(x) = x^2$ with solution $\phi(t, \xi) = \xi/(1 - \xi t)$. Consider the compact interval $K := [-r, r] \subset \mathbb{R}$, $r > 0$. Clearly $\phi(h, \xi)$ with $\xi = r$ and $h = 1/r$ is not defined (division by zero) and therefore the estimate (3) cannot hold true for all $\xi \in K$ and and $h > 0$.

To circumvent this, fix an $R > r$. For solutions starting in K and not leaving $[-R, R] \supset K$ for times in $[0, t]$ we have

$$|\phi(t, \xi)| \leq |\phi(0, \xi)| + \int_0^t \max_{x \in [-R, R]} |f(x)| dt \leq r + tR^2.$$

Hence, a solution starting in K will stay in $[-R, R]$ at least for $0 \leq t \leq (R - r)/R^2$, because then $|\phi(t, \xi)| \leq R$. It follows that we can choose $h^* = (R-r)/R^2$ and for all $0 < h \leq h^*$ the estimate (3) will hold true for an appropriate constant C_{RK4}. □

Runge-Kutta methods exist of any order, i.e. for any $p \in \mathbb{N}$ one can derive formulas similar to (2) and obtain constants $h^* > 0$ and C_{RKp} such that

$$\|\widetilde{\boldsymbol{\phi}}_1(\boldsymbol{\xi}) - \boldsymbol{\phi}(h, \boldsymbol{\xi})\| \leq C_{\mathrm{RKp}} h^{p+1} \tag{4}$$

for all $\boldsymbol{\xi} \in K$ and all $0 < h \leq h^*$. The constant C_{RKp} depends on up to the p-th order derivatives of $\mathbf{f}$ in some compact and convex set $\widetilde{K} \supset K$, which is sufficiently large. We refer to a Runge-Kutta method of order $p \in \mathbb{N}$ as RKp.

2.2 Multi-Step Methods

In a p-step method, $p \in \mathbb{N}$, $\widetilde{\boldsymbol{\phi}}_{i+1}$ is computed from $\widetilde{\boldsymbol{\phi}}_j$ with $j = i, i-1, \ldots, i-p+1$. Obviously a one-step method has $p = 1$, and if $p > 1$ then the method is called a multi-step method. A typical example is the Adams-Bashforth four-step method (AB4), given by:

Fix the step-size $h > 0$ and set $\widetilde{\boldsymbol{\phi}}_0 = \boldsymbol{\xi}$, $\widetilde{\boldsymbol{\phi}}_j \approx \boldsymbol{\phi}(jh, \boldsymbol{\xi})$ for $j = 1, 2, 3$. Then, for every $i \in \mathbb{N}$, $i \geq 3$, set

$$\widetilde{\boldsymbol{\phi}}_{i+1} = \widetilde{\boldsymbol{\phi}}_i + \frac{h}{24}\left(55\mathbf{f}(\widetilde{\boldsymbol{\phi}}_i) - 59\mathbf{f}(\widetilde{\boldsymbol{\phi}}_{i-1}) + 37\mathbf{f}(\widetilde{\boldsymbol{\phi}}_{i-2}) - 9\mathbf{f}(\widetilde{\boldsymbol{\phi}}_{i-3})\right). \tag{5}$$

For multi-step methods like AB4 the error estimate is usually formulated differently to one-step methods: for a compact set $K \subset \mathbb{R}^n$ there exist constants $h^* > 0$ and $C_{\mathrm{AB4}} > 0$ such that for all $0 < h \leq h^*$ we have

$$\|\widetilde{\boldsymbol{\phi}}_{i+1}(\boldsymbol{\xi}) - \boldsymbol{\phi}((i+1)h, \boldsymbol{\xi})\| \leq C_{\mathrm{AB4}} h^5 \tag{6}$$

if $\widetilde{\boldsymbol{\phi}}_j = \boldsymbol{\phi}(jh, \boldsymbol{\xi}) \in K$ for $j = i, i-1, i-2, i-3$ in formula (5), i.e. if the previous approximations $\widetilde{\boldsymbol{\phi}}_j$ are exact and are all in K. Again, the constant C_{AB4} depends on the derivatives of $\mathbf{f}$, up to and including the fourth order, in a compact and convex set $\widetilde{K} \supset K$ that is sufficiently large.

Since the values $\boldsymbol{\phi}(jh, \boldsymbol{\xi})$ for $j = i, i-1, i-2, i-3$ are usually not known, we cannot set $\widetilde{\boldsymbol{\phi}}_j = \boldsymbol{\phi}(jh, \boldsymbol{\xi})$ for $j = i, i-1, i-2, i-3$. These values are usually fixed by using a one-step method for the first three steps, e.g. by using RK4.

For any $p \in \mathbb{N}$ one can define the p-step Adams-Bashforth method (ABp). The general formula is

$$\widetilde{\boldsymbol{\phi}}_{i+1} = \widetilde{\boldsymbol{\phi}}_i + h \sum_{j=0}^{p-1} a_j \mathbf{f}(\widetilde{\boldsymbol{\phi}}_{i-j}), \quad \text{where} \quad \sum_{j=0}^{p-1} a_j = 1. \tag{7}$$

Again, for a compact set $K \subset \mathbb{R}^n$ and a maximum step-size $h^* > 0$, there exists a constant $C_{\mathrm{ABp}} > 0$ such that

$$\|\widetilde{\boldsymbol{\phi}}_{i+1}(\boldsymbol{\xi}) - \boldsymbol{\phi}((i+1)h, \boldsymbol{\xi})\| \leq C_{\mathrm{ABp}} h^{p+1} \tag{8}$$

for all step-sizes $0 < h \leq h^*$, if $\widetilde{\boldsymbol{\phi}}_j = \boldsymbol{\phi}(jh, \boldsymbol{\xi}) \in K$ for $j = i, i-1, \ldots, i-p+1$ in formula (7), i.e. if the previous approximations $\widetilde{\boldsymbol{\phi}}_j$ are exact and are all in K. Unsurprisingly, the constant C_{ABp} depends on the derivatives of $\mathbf{f}$, up to and including the p-th order, in a compact and convex set $\widetilde{K} \supset K$ that is sufficiently large. The first values $\widetilde{\boldsymbol{\phi}}_j(\boldsymbol{\xi})$, $j = 1, 2, \ldots, p-1$ are usually initialized by using a p-th order one-step method, e.g. RKp.

2.3 Corrector Step

Commonly, when the Adams-Bashforth method is used to solve an initial-value problem, a so-called corrector step is added. That is, the value $\widetilde{\boldsymbol{\phi}}_{i+1}$ is updated after the initial computation using the implicit Adams-Moulton (AM) method. To give a concrete example, consider the AB4 method above, but set

$$\widetilde{\boldsymbol{\phi}}_{i+1}^{\mathrm{pre}} = \widetilde{\boldsymbol{\phi}}_i + \frac{h}{24}\left(55\mathbf{f}(\widetilde{\boldsymbol{\phi}}_i) - 59\mathbf{f}(\widetilde{\boldsymbol{\phi}}_{i-1}) + 37\mathbf{f}(\widetilde{\boldsymbol{\phi}}_{i-2}) - 9\mathbf{f}(\widetilde{\boldsymbol{\phi}}_{i-3})\right). \tag{9}$$

to underline that we are going to update the *predicted value* $\widetilde{\boldsymbol{\phi}}_{i+1}^{\mathrm{pre}}$ later. The corrected value is then computed using the three-step Adams-Moulton (AM4) method, i.e.

$$\widetilde{\boldsymbol{\phi}}_{i+1} = \widetilde{\boldsymbol{\phi}}_i + \frac{h}{24}\left(9\mathbf{f}(\widetilde{\boldsymbol{\phi}}_{i+1}^{\mathrm{pre}}) + 19\mathbf{f}(\widetilde{\boldsymbol{\phi}}_i) - 5\mathbf{f}(\widetilde{\boldsymbol{\phi}}_{i-1}) + \mathbf{f}(\widetilde{\boldsymbol{\phi}}_{i-2})\right). \tag{10}$$

We denote this combined method by PC4, i.e. first *predicting* using AB4 and then *correcting* using AM4; we call it AM4, and not AM3, although it only uses the three last steps, because it can be shown to be a fourth-order method. The error estimate can be given similarly to above for the AB4 method. That is, for a compact set $K \subset \mathbb{R}^n$ there exists constants $h^* > 0$ and $C_{\mathrm{PC4}} > 0$ such that for all $0 < h \leq h^*$ we have

$$\|\widetilde{\boldsymbol{\phi}}_{i+1}(\boldsymbol{\xi}) - \boldsymbol{\phi}((i+1)h, \boldsymbol{\xi})\| \leq C_{\mathrm{PC4}} h^5 \tag{11}$$

if $\widetilde{\boldsymbol{\phi}}_j = \boldsymbol{\phi}(jh, \boldsymbol{\xi}) \in K$ for $j = i, i-1, i-2, i-3$ in formula (10) and (9), i.e. if the previous approximations $\widetilde{\boldsymbol{\phi}}_j$ are exact and are all in K. Once again, the constant C_{PC4} depends on the derivatives of $\mathbf{f}$, up to and including the fourth order, in a compact and convex set $\widetilde{K} \supset K$ that is sufficiently large. As before, the values $\widetilde{\boldsymbol{\phi}}_j$ for $j = i, i-1, i-2, i-3$ are in practice computed using a one-step method, e.g. RK4.

Again, one can for any $k \in \mathbb{N}$ define PCp using ABp for the prediction and AMp for the correction. The general formula is then given by

$$\widetilde{\boldsymbol{\phi}}_{i+1} = \widetilde{\boldsymbol{\phi}}_i + h\left(b_{-1}\mathbf{f}(\widetilde{\boldsymbol{\phi}}^{\mathrm{pre}}_{i+1}) + \sum_{j=0}^{p-2} b_j \mathbf{f}(\widetilde{\boldsymbol{\phi}}_{i-j})\right), \quad \text{where} \quad \sum_{j=-1}^{p-2} b_j = 1 \tag{12}$$

and $\widetilde{\boldsymbol{\phi}}^{\mathrm{pre}}_{i+1}$ is equal to the right-hand-side of (7), i.e.

$$\widetilde{\boldsymbol{\phi}}^{\mathrm{pre}}_{i+1} = \widetilde{\boldsymbol{\phi}}_i + h\sum_{j=0}^{p-1} a_j \mathbf{f}(\widetilde{\boldsymbol{\phi}}_j), \quad \text{where} \quad \sum_{j=0}^{p-1} a_j = 1. \tag{13}$$

As before, there exist constants $h^* > 0$ and $C_{\mathrm{PCp}} > 0$ for each compact set $K \subset \mathbb{R}^n$, such that

$$\|\widetilde{\boldsymbol{\phi}}_{i+1}(\boldsymbol{\xi}) - \boldsymbol{\phi}((i+1)h, \boldsymbol{\xi})\| \le C_{\mathrm{PCp}} h^{p+1} \tag{14}$$

for all step-sizes $0 < h \le h^*$, if $\widetilde{\boldsymbol{\phi}}_j = \boldsymbol{\phi}(jh, \boldsymbol{\xi}) \in K$ for $j = i, i-1, \ldots, i-p+1$ in formulas (13) and (12), i.e. if the previous approximations $\widetilde{\boldsymbol{\phi}}_j$ are exact and are all in K. The constant C_{PCp} depends on the derivatives of $\mathbf{f}$, up to and including the p-th order, in a compact and convex set $\widetilde{K} \supset K$ that is sufficiently large. The first values $\widetilde{\boldsymbol{\phi}}_j(\boldsymbol{\xi})$, $j = 1, 2, \ldots, p-1$ are usually initialized as in the ABp method by using a p-th order one-step method, e.g. RKp.

2.4 Uniform Error Estimate

The error estimates (8) and (14) (also (6) and (11)) are not in a form which is useful for our application of computing Lyapunov functions and contraction metrics. We need require error estimates as in the following definition, uniform for all the methods, where one does not assume $\widetilde{\boldsymbol{\phi}}_j(\boldsymbol{\xi}) = \boldsymbol{\phi}(jh, \boldsymbol{\xi})$ for the previous steps.

Definition 1. *(Order of numerical methods)*
A numerical method to solve

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

is said to be of order $p \in \mathbb{N}$, if for any compact set $S \subset \mathbb{R}^n$ there exist constants $C, h^* > 0$ such that for all step-sizes $0 < h \le h^*$ we have for any $i \in \mathbb{N}_0$ that

$$\|\widetilde{\boldsymbol{\phi}}_{i+1}(\boldsymbol{\xi}) - \boldsymbol{\phi}(h, \widetilde{\boldsymbol{\phi}}_i(\boldsymbol{\xi}))\| \le Ch^{p+1}$$

whenever

$$\widetilde{\phi_0}(\boldsymbol{\xi}), \widetilde{\phi}_1(\boldsymbol{\xi}), \ldots, \widetilde{\phi}_i(\boldsymbol{\xi}) \in S.$$

We now prove that all the numerical integration methods we have discussed, that is,

(i) RKp,
(ii) ABp initialized with RKp,
(iii) PCp initialized with RKp,

are all of order p in the sense of Definition 1.

Theorem 1. *(Error estimate for RKp, ABp, and PCp) Consider the system* (1) *and assume that* $\mathbf{f} \in C^p(\mathbb{R}^n; \mathbb{R}^n)$*. Then the numerical integration methods (i) RKp, (ii) ABp initialized with RKp, and (iii) PCp initialized with RKp, are of order* $p \in \mathbb{N}$ *in the sense of Definition 1.*

Proof. The case (i) is simple, as it follows directly from the discussion in Section 2.1. Note that $p = 1$ corresponds to a single step method and reduces to case (i). For the rest of the proof let the order $p \in \mathbb{N}$, $p \geq 2$, be fixed and $0 < h_0 \leq 1$ be small enough, so that (4) holds for all $\boldsymbol{\xi} \in S$ and all $0 < h \leq h_0$. We prove the cases (ii) and (iii) successively.

Case (ii): ABp initialized by RKp.

We will first define the sets $S, S', \widetilde{S}$ with the following goal: S' contains the next iterate $(i+1)$ of both the numerical approximation and the true solution, if the first i iterates lie in S; $\widetilde{S}$ is an even larger set that ensures that if we apply the backward flow $\boldsymbol{\phi}([-(p-1)h_2, 0], \mathbf{y})$ to an element $\mathbf{y}$ of S', then the result is contained in $\widetilde{S}$. Moreover, we choose the step-size $h_2 > 0$ sufficiently small that the backward solution exists.

In detail, we fix a constant $0 < h_1 \leq h_0$ and a compact, convex set $S' \supset S$ such that $\boldsymbol{\phi}(h, \boldsymbol{\xi}) \in S'$ for all $\boldsymbol{\xi} \in S$ and $\widetilde{\boldsymbol{\phi}}_{i+1}(\boldsymbol{\xi}) \in S'$, whenever $\widetilde{\boldsymbol{\phi}}_j(\boldsymbol{\xi}) \in S$, for $j = 0, 1, 2, \ldots, i$, $i \in \mathbb{N}_0$, and when using step-size $0 < h \leq h_1$. Furthermore, let $0 < h_2 \leq h_1$ be a constant and $\widetilde{S}$ be a compact, convex set such that $\boldsymbol{\phi}([-(p-1)h_2, 0], S')$ exists and

$$\boldsymbol{\phi}([-(p-1)h_2, 0], S') \subset \widetilde{S}.$$

Fix an arbitrary, but constant $\boldsymbol{\xi} \in S$ for the rest of the proof of case (ii).

For a fixed step-size $0 < h \leq h_2$ denote $\boldsymbol{\phi}_i(\mathbf{y}) := \boldsymbol{\phi}(ih, \mathbf{y})$, $\mathbf{y} \in S$ and $i \geq -(p-1)$ with $i \in \mathbb{Z}$, which is defined by the above assumptions. Furthermore, denote by $\widetilde{\boldsymbol{\phi}}_i$, $i \in \mathbb{N}_0$, the approximation to $\boldsymbol{\phi}_i(\boldsymbol{\xi})$ generated by ABp initialized by RKp.

For the steps with the ABp method let us define, with $a_0, a_1, \ldots, a_{p-1}$ the constants from (7),

$$\mathrm{ABp}(\mathbf{x}_i, \mathbf{x}_{i-1}, \ldots, \mathbf{x}_{i-(p-1)}) := \mathbf{x}_i + h \sum_{j=0}^{p-1} a_j \mathbf{f}(\mathbf{x}_{i-j}), \tag{15}$$

We consider two choices of the $\mathbf{x}_j$; if we choose them to be the iterates of the numerical method we obtain by the formula (7) for the ABp method that

$$\widetilde{\boldsymbol{\phi}}_{i+1} = \mathrm{ABp}(\widetilde{\boldsymbol{\phi}}_i, \ldots, \widetilde{\boldsymbol{\phi}}_{i-(p-1)}) \quad \text{if} \quad i \geq p-1.$$

If we choose them to be the true solution, then we have the following result: if $0 < h \leq h_2$ and $\mathbf{y}$ satisfies $\boldsymbol{\phi}_i(\mathbf{y}) \in S'$, then $\boldsymbol{\phi}_j(\mathbf{y}) \in \widetilde{S}$ for $j = i, i-1, \ldots, i-(p-1)$ by the definition of $\widetilde{S}$ and there exists a constant $C_{\mathrm{ABp}} > 0$, such that

$$\| \mathrm{ABp}(\boldsymbol{\phi}_i(\mathbf{y}), \ldots, \boldsymbol{\phi}_{i-(p-1)}(\mathbf{y})) - \boldsymbol{\phi}_{i+1}(\mathbf{y}) \| \leq C_{\mathrm{ABp}} h^{p+1}. \tag{16}$$

This follows from (8) by setting $K := \widetilde{S}$.

We now prove case (ii) by induction.

Denote by (I) the proposition:

> There exist constants $C, C^*, h^* > 0$, such that for every time-step $0 < h \leq h^*$ we have for any $i \in \mathbb{N}$ that
>
> $$\|\widetilde{\boldsymbol{\phi}}_i - \boldsymbol{\phi}_1(\widetilde{\boldsymbol{\phi}}_{i-1})\| \leq C h^{p+1}, \tag{17}$$
>
> whenever $\widetilde{\boldsymbol{\phi}}_k \in S$ for $k = 0, 1, \ldots, i-1$, and additionally we have for $i \geq p-1$ and with $j = 0, 1, \ldots, p-1$, that
>
> $$\|\widetilde{\boldsymbol{\phi}}_{i-j} - \boldsymbol{\phi}_{-j}(\widetilde{\boldsymbol{\phi}}_i)\| \leq C^* h^{p+1}. \tag{18}$$

Recall that $\boldsymbol{\phi}_{-j}(\widetilde{\boldsymbol{\phi}}_i)$ is defined by the assumptions and lies in $\widetilde{S}$. Note that case (ii) follows from (17) in (I); (18) is just needed for the induction.

To prove (I) let us first fix the constants. Set

$$A := \sum_{j=1}^{p-1} |a_j|,$$

let $L > 0$ be a Lipschitz constant for $\mathbf{f}$ on $\widetilde{S}$, i.e. $\|\mathbf{f}(\mathbf{x}) - \mathbf{f}(\mathbf{y})\| \leq L\|\mathbf{x} - \mathbf{y}\|$ for all $\mathbf{x}, \mathbf{y} \in \widetilde{S}$, and set

$$C := 2\max\{C_{\mathrm{RKp}}, C_{\mathrm{ABp}}\},$$

$$C^* := C \sum_{k=1}^{p-1} e^{kL},$$

and

$$h^* := \min\left\{h_2, \frac{C - C_{\mathrm{ABp}}}{ALC^*}\right\} > 0. \tag{19}$$

We first note that case (i) implies that (I) holds true for $i = 0, 1, \ldots, p-1$, i.e. when we are using RKp to generate the values $\widetilde{\boldsymbol{\phi}}_i$. Indeed, (17) follows directly from case (i).

To show (18) for $i = p-1$ and $j = 0, 1, \ldots, p-1$, we have, noting that $\widetilde{\boldsymbol{\phi}}_{i-j+k-1} \in \widetilde{S}$ by assumption,

$$\begin{aligned}
\|\widetilde{\boldsymbol{\phi}}_{i-j} - \boldsymbol{\phi}_{-j}(\widetilde{\boldsymbol{\phi}}_i)\| &= \|\boldsymbol{\phi}_0(\widetilde{\boldsymbol{\phi}}_{i-j}) - \boldsymbol{\phi}_{-j}(\widetilde{\boldsymbol{\phi}}_i)\| \\
&= \left\| \sum_{k=1}^{j} \left[\boldsymbol{\phi}_{-k+1}(\widetilde{\boldsymbol{\phi}}_{i-j+k-1}) - \boldsymbol{\phi}_{-k}(\widetilde{\boldsymbol{\phi}}_{i-j+k}) \right] \right\| \\
&\leq \sum_{k=1}^{j} \|\boldsymbol{\phi}_{-k+1}(\widetilde{\boldsymbol{\phi}}_{i-j+k-1}) - \boldsymbol{\phi}_{-k}(\widetilde{\boldsymbol{\phi}}_{i-j+k})\| \\
&\leq \sum_{k=1}^{j} e^{khL} \|\boldsymbol{\phi}_1(\widetilde{\boldsymbol{\phi}}_{i-j+k-1}) - \boldsymbol{\phi}_0(\widetilde{\boldsymbol{\phi}}_{i-j+k})\| \\
&\leq \sum_{k=1}^{j} e^{khL} \|\boldsymbol{\phi}_1(\widetilde{\boldsymbol{\phi}}_{i-j+k-1}) - \widetilde{\boldsymbol{\phi}}_{1+(i-j+k-1)}\| \\
&\leq \sum_{k=1}^{j} e^{khL} C_{\mathrm{RKp}} h^{p+1} \\
&\leq C \sum_{k=1}^{p-1} e^{kL} h^{p+1} \\
&= C^* h^{p+1}.
\end{aligned}$$

where we have used the well known estimate

$$\|\boldsymbol{\phi}(t, \mathbf{a}) - \boldsymbol{\phi}(t, \mathbf{b})\| \leq e^{L|t|} \|\mathbf{a} - \mathbf{b}\| \tag{20}$$

as well as (4) with $K = \widetilde{S}$.

We now show that (I) holds true for $i \geq p$. For this assume that (I) holds true for all natural numbers up to and including some $i \geq p-1$. We assume that $\widetilde{\boldsymbol{\phi}}_k \in S$ for $k = 0, \ldots, i$ and show that (I) also holds true for $i+1$.

Let us first consider (17) with i replaced by $i+1$. Observe that

$$\begin{aligned}
\|\widetilde{\boldsymbol{\phi}}_{i+1} - \boldsymbol{\phi}_1(\widetilde{\boldsymbol{\phi}}_i)\| &= \|\,\mathrm{ABp}(\widetilde{\boldsymbol{\phi}}_i, \ldots, \widetilde{\boldsymbol{\phi}}_{i-(p-1)}) - \boldsymbol{\phi}_1(\widetilde{\boldsymbol{\phi}}_i)\| \\
&\leq \|\,\mathrm{ABp}(\widetilde{\boldsymbol{\phi}}_i, \ldots, \widetilde{\boldsymbol{\phi}}_{i-(p-1)}) - \mathrm{ABp}(\boldsymbol{\phi}_0(\widetilde{\boldsymbol{\phi}}_i), \ldots, \boldsymbol{\phi}_{-(p-1)}(\widetilde{\boldsymbol{\phi}}_i))\| \\
&\quad + \|\,\mathrm{ABp}(\boldsymbol{\phi}_0(\widetilde{\boldsymbol{\phi}}_i), \ldots, \boldsymbol{\phi}_{-(p-1)}(\widetilde{\boldsymbol{\phi}}_i)) - \boldsymbol{\phi}_1(\widetilde{\boldsymbol{\phi}}_i)\|
\end{aligned} \tag{21}$$

and for the second term on the right-hand-side we have the bound

$$\|\,\mathrm{ABp}(\boldsymbol{\phi}_0(\widetilde{\boldsymbol{\phi}}_i), \ldots, \boldsymbol{\phi}_{-(p-1)}(\widetilde{\boldsymbol{\phi}}_i)) - \boldsymbol{\phi}_1(\widetilde{\boldsymbol{\phi}}_i)\| \leq C_{\mathrm{ABp}} h^{p+1} \tag{22}$$

by (16) since $\widetilde{\boldsymbol{\phi}}_i \in S \subset S'$.

To bound the first term on the right-hand-side of (21) we use the formula for ABp, $\boldsymbol{\phi}_0(\widetilde{\boldsymbol{\phi}}_i) = \widetilde{\boldsymbol{\phi}}_i$, the Lipschitz condition on $\mathbf{f}$ on $\widetilde{S}$, and the induction hypothesis (18), and we get

$$\begin{aligned}
\| \mathrm{ABp}(\widetilde{\boldsymbol{\phi}}_i, \ldots, \widetilde{\boldsymbol{\phi}}_{i-(p-1)}) &- \mathrm{ABp}(\boldsymbol{\phi}_0(\widetilde{\boldsymbol{\phi}}_i), \ldots, \boldsymbol{\phi}_{-(p-1)}(\widetilde{\boldsymbol{\phi}}_i))\| \\
&\leq h \left\| \sum_{j=0}^{p-1} a_j \left[\mathbf{f}(\widetilde{\boldsymbol{\phi}}_{i-j}) - \mathbf{f}(\boldsymbol{\phi}_{-j}(\widetilde{\boldsymbol{\phi}}_i)) \right] \right\| \\
&\leq h \sum_{j=1}^{p-1} |a_j| \left\| \mathbf{f}(\widetilde{\boldsymbol{\phi}}_{i-j}) - \mathbf{f}(\boldsymbol{\phi}_{-j}(\widetilde{\boldsymbol{\phi}}_i)) \right\| \\
&\leq h \sum_{j=1}^{p-1} |a_j| L \left\| \widetilde{\boldsymbol{\phi}}_{i-j} - \boldsymbol{\phi}_{-j}(\widetilde{\boldsymbol{\phi}}_i) \right\| \\
&\leq h \sum_{j=1}^{p-1} |a_j| L C^* h^{p+1} \\
&= A L C^* h^{p+2}. \qquad (23)
\end{aligned}$$

Note that $\widetilde{\boldsymbol{\phi}}_{i-j} \in S \subset \widetilde{S}$ and $\boldsymbol{\phi}_{-j}(\widetilde{\boldsymbol{\phi}}_i) \in \widetilde{S}$. Hence, (21), (22), and (23) deliver

$$\|\widetilde{\boldsymbol{\phi}}_{i+1} - \boldsymbol{\phi}_1(\widetilde{\boldsymbol{\phi}}_i)\| \leq A L C^* h^{p+2} + C_{\mathrm{ABp}} h^{p+1} \leq C h^{p+1}, \qquad (24)$$

because

$$A L C^* h + C_{\mathrm{ABp}} \leq A L C^* h^* + C_{\mathrm{ABp}} \leq C$$

by (19). Hence, the bound (17) in (I) holds true for i replaced by $i+1$.

Let us now consider the bound (18) in (I) for i replaced by $i+1$. The case $j = 0$ is obvious and from

$$\begin{aligned}
\|\widetilde{\boldsymbol{\phi}}_{i+1-j} - \boldsymbol{\phi}_{-j}(\widetilde{\boldsymbol{\phi}}_{i+1})\| &= \|\boldsymbol{\phi}_{-j}(\boldsymbol{\phi}_j(\widetilde{\boldsymbol{\phi}}_{i+1-j})) - \boldsymbol{\phi}_{-j}(\widetilde{\boldsymbol{\phi}}_{i+1})\| \\
&\leq e^{jLh} \|\boldsymbol{\phi}_j(\widetilde{\boldsymbol{\phi}}_{i+1-j}) - \widetilde{\boldsymbol{\phi}}_{i+1}\|
\end{aligned}$$

the case $j = 1$, i.e.

$$\|\widetilde{\boldsymbol{\phi}}_i - \boldsymbol{\phi}_{-1}(\widetilde{\boldsymbol{\phi}}_{i+1})\| \leq e^{Lh} \|\boldsymbol{\phi}_1(\widetilde{\boldsymbol{\phi}}_i) - \widetilde{\boldsymbol{\phi}}_{i+1}\| \leq C^* h^{p+1}$$

follows from (24), $e^{Lh}C \le C^*$, and (20). The general case for $j \le p-1$ follows similarly from (24), the induction hypothesis (17), and the definition of C^*:

$$\begin{aligned}
\|\widetilde{\boldsymbol{\phi}}_{i+1-j} - \boldsymbol{\phi}_{-j}(\widetilde{\boldsymbol{\phi}}_{i+1})\| &= \|\boldsymbol{\phi}_0(\widetilde{\boldsymbol{\phi}}_{i+1-j}) - \boldsymbol{\phi}_{-j}(\widetilde{\boldsymbol{\phi}}_{i+1})\| \\
&= \left\| \sum_{k=1}^{j} \left[\boldsymbol{\phi}_{1-k}(\widetilde{\boldsymbol{\phi}}_{i-j+k}) - \boldsymbol{\phi}_{1-(k+1)}(\widetilde{\boldsymbol{\phi}}_{i-j+k+1}) \right] \right\| \\
&\le \sum_{k=1}^{j} \|\boldsymbol{\phi}_{1-k}(\widetilde{\boldsymbol{\phi}}_{i-j+k}) - \boldsymbol{\phi}_{-k}(\widetilde{\boldsymbol{\phi}}_{i-j+k+1})\| \\
&\le \sum_{k=1}^{j} e^{khL} \|\boldsymbol{\phi}_1(\widetilde{\boldsymbol{\phi}}_{i-j+k}) - \boldsymbol{\phi}_0(\widetilde{\boldsymbol{\phi}}_{i-j+k+1})\| \\
&\le \sum_{k=1}^{j} e^{khL} \|\boldsymbol{\phi}_1(\widetilde{\boldsymbol{\phi}}_{i-j+k}) - \widetilde{\boldsymbol{\phi}}_{1+i-j+k}\| \\
&\le \sum_{k=1}^{j} e^{khL} C h^{p+1} \\
&\le \sum_{k=1}^{p-1} e^{kL} C h^{p+1} \\
&= C^* h^{p+1}.
\end{aligned}$$

Thus, we have proved the bound (18) of (I) for i replaced by $i+1$, which shows case (ii).

Case (iii): PCp initialized by RKp.

As in case (ii), fix a constant $0 < h_1 \le h_0$ and a compact, convex set $S' \supset S$ such that $\boldsymbol{\phi}(h, \boldsymbol{\xi}) \in S'$ for all $\boldsymbol{\xi} \in S$ and $\widetilde{\boldsymbol{\phi}}^{\text{pre}}_{i+1}(\boldsymbol{\xi}), \widetilde{\boldsymbol{\phi}}_{i+1}(\boldsymbol{\xi}) \in S'$, whenever $\widetilde{\boldsymbol{\phi}}_j(\boldsymbol{\xi}) \in S$, for $j = 0, 1, 2, \ldots, i$, $i \in \mathbb{N}_0$, and when using step-size $0 < h \le h_1$. Furthermore, let $0 < h_2 \le h_1$ be a constant and $\widetilde{S}$ be a compact, convex set such that $\phi([-(p-1)h_2, 0], S')$ exists and $\phi([-(p-1)h_2, 0], S') \subset \widetilde{S}$.

Fix an arbitrary, but constant $\boldsymbol{\xi} \in S$ for the rest of the proof.

For a fixed step-size $0 < h \le h_2$ denote $\boldsymbol{\phi}_i(\mathbf{y}) := \boldsymbol{\phi}(ih, \mathbf{y})$, $\mathbf{y} \in S$ and $i \ge -(p-1)$ with $i \in \mathbb{Z}$, which is defined by the above assumptions. Furthermore, denote by $\widetilde{\boldsymbol{\phi}}_i$, $i \in \mathbb{N}_0$, the approximation to $\boldsymbol{\phi}_i(\boldsymbol{\xi})$ generated by PCp initialized by RKp. That is, $\widetilde{\boldsymbol{\phi}}_0 = \boldsymbol{\xi}$, the values $\widetilde{\boldsymbol{\phi}}_i$ are generated by RKp for $i = 1, 2, \ldots, p-1$ and for $i = p, p+1, \ldots$ we set, using (12)

$$\widetilde{\boldsymbol{\phi}}_{i+1} = \mathrm{PCp}(\widetilde{\boldsymbol{\phi}}_i, \widetilde{\boldsymbol{\phi}}_{i-1}, \ldots, \widetilde{\boldsymbol{\phi}}_{i-(p-1)}),$$

where

$$\mathrm{PCp}(\mathbf{x}_i,\dots,\mathbf{x}_{i-(p-1)}) := \mathbf{x}_i + h\left(b_{-1}\mathbf{f}(\mathrm{ABp}(\mathbf{x}_i,\dots,\mathbf{x}_{i-(p-1)})) + \sum_{j=0}^{p-2} b_j\mathbf{f}(\mathbf{x}_{i-j})\right)$$

is defined similarly to ABp in (15), but using the AMp formula (12).

If $0 < h \le h_2$ and $\mathbf{y}$ satisfies $\boldsymbol{\phi}_i(\mathbf{y}) \in S'$, then $\boldsymbol{\phi}_j(\mathbf{y}) \in \widetilde{S}$ for $j = i, i-1, \dots, i-(p-1)$ by definition of $\widetilde{S}$ and there exists a constant $C_{\mathrm{PCp}} > 0$, such that

$$\|\,\mathrm{PCp}(\boldsymbol{\phi}_i(\mathbf{y}),\dots,\boldsymbol{\phi}_{i-(p-1)}(\mathbf{y})) - \boldsymbol{\phi}_{i+1}(\mathbf{y})\| \le C_{\mathrm{PCp}} h^{p+1}. \tag{25}$$

This follows from (14) by setting $K := \widetilde{S}$.

We now prove case (iii) by induction, similar to how we proved case (ii).

Denote by (II) the proposition:

> There exist constants $C, C^*, h^* > 0$, such that for every time-step $0 < h \le h^*$ we have for any $i \in \mathbb{N}_0$ that
>
> $$\|\widetilde{\boldsymbol{\phi}}_i - \boldsymbol{\phi}_1(\widetilde{\boldsymbol{\phi}}_{i-1})\| \le C h^{p+1}, \tag{26}$$
>
> whenever $\widetilde{\boldsymbol{\phi}}_k \in S$ for $k = 0, 1, \dots, i-1$, and additionally we have for $i \ge p-1$ and with $j = 0, 1, \dots, p-1$, that
>
> $$\|\widetilde{\boldsymbol{\phi}}_{i-j} - \boldsymbol{\phi}_{-j}(\widetilde{\boldsymbol{\phi}}_i)\| \le C^* h^{p+1}. \tag{27}$$

Note that case (iii) follows from (26) in (II); (27) is just needed for the induction.

To prove (II) let us first fix the constants. Set

$$A := \sum_{j=1}^{p-1} |a_j|, \quad B := \sum_{j=1}^{p-2} |b_j|,$$

let $L > 0$ be a Lipschitz constant for $\mathbf{f}$ on $\widetilde{S}$ and set

$$C := 2\max\{C_{\mathrm{RKp}}, C_{\mathrm{PCp}}\},$$

$$C^* := C\sum_{k=1}^{p-1} e^{kL},$$

and

$$h^* := \min\left\{h_2, \frac{C - C_{\mathrm{PCp}}}{LC^*(B + AL|b_{-1}|)}\right\} > 0. \tag{28}$$

As in the proof of case (ii) we see that (II) holds true for $i = 0, 1, \dots, p-1$, i.e. when we are using RKp to generate the values $\widetilde{\boldsymbol{\phi}}_i$.

We now show that (II) holds true for $i \ge p$. For this assume that (II) holds true for all natural numbers up to and including some $i \ge p-1$. We assume that $\widetilde{\boldsymbol{\phi}}_k \in S$ for all $k = 0, \dots, i$ and show that (II) also holds true for $i+1$.

Let us first consider (26) with i replaced by $i+1$. Observe that

$$\begin{aligned}\|\widetilde{\boldsymbol{\phi}}_{i+1} - \boldsymbol{\phi}_1(\widetilde{\boldsymbol{\phi}}_i)\| &= \|\operatorname{PCp}(\widetilde{\boldsymbol{\phi}}_i, \ldots, \widetilde{\boldsymbol{\phi}}_{i-(p-1)}) - \boldsymbol{\phi}_1(\widetilde{\boldsymbol{\phi}}_i)\| \\ &\le \|\operatorname{PCp}(\widetilde{\boldsymbol{\phi}}_i, \ldots, \widetilde{\boldsymbol{\phi}}_{i-(p-1)}) - \operatorname{PCp}(\boldsymbol{\phi}_0(\widetilde{\boldsymbol{\phi}}_i), \ldots, \boldsymbol{\phi}_{-(p-1)}(\widetilde{\boldsymbol{\phi}}_i))\| \\ &+ \|\operatorname{PCp}(\boldsymbol{\phi}_0(\widetilde{\boldsymbol{\phi}}_i), \ldots, \boldsymbol{\phi}_{-(p-1)}(\widetilde{\boldsymbol{\phi}}_i)) - \boldsymbol{\phi}_1(\widetilde{\boldsymbol{\phi}}_i)\| \end{aligned} \tag{29}$$

and for the second term on the right-hand-side we have the bound

$$\|\operatorname{PCp}(\boldsymbol{\phi}_0(\widetilde{\boldsymbol{\phi}}_i), \ldots, \boldsymbol{\phi}_{-(p-1)}(\widetilde{\boldsymbol{\phi}}_i)) - \boldsymbol{\phi}_1(\widetilde{\boldsymbol{\phi}}_i)\| \le C_{\mathrm{PCp}} h^{p+1} \tag{30}$$

by (25).

To bound the first term on the right-hand-side of (29) we use the formula for PCp, that $\boldsymbol{\phi}_0(\widetilde{\boldsymbol{\phi}}_i) = \widetilde{\boldsymbol{\phi}}_i$, the Lipschitz condition on $\mathbf{f}$ on $\widetilde{S}$, and induction hypothesis (27), and we get

$$\begin{aligned}&\|\operatorname{PCp}(\widetilde{\boldsymbol{\phi}}_i, \ldots, \widetilde{\boldsymbol{\phi}}_{i-(p-1)}) - \operatorname{PCp}(\boldsymbol{\phi}_0(\widetilde{\boldsymbol{\phi}}_i), \ldots, \boldsymbol{\phi}_{-(p-1)}(\widetilde{\boldsymbol{\phi}}_i))\| \\ &\quad\le h\Bigg\| b_{-1}\Big(\mathbf{f}(\operatorname{ABp}(\widetilde{\boldsymbol{\phi}}_i, \ldots, \widetilde{\boldsymbol{\phi}}_{i-(p-1)})) - \mathbf{f}(\operatorname{ABp}(\boldsymbol{\phi}_0(\widetilde{\boldsymbol{\phi}}_i), \ldots, \boldsymbol{\phi}_{-(p-1)}(\widetilde{\boldsymbol{\phi}}_i))\Big) \\ &\qquad + \sum_{j=0}^{p-2} b_j \Big[\mathbf{f}(\widetilde{\boldsymbol{\phi}}_{i-j}) - \mathbf{f}(\boldsymbol{\phi}_{-j}(\widetilde{\boldsymbol{\phi}}_i))\Big] \Bigg\| \\ &\quad\le hL|b_{-1}| \left\|\operatorname{ABp}(\widetilde{\boldsymbol{\phi}}_i, \ldots, \widetilde{\boldsymbol{\phi}}_{i-(p-1)}) - \operatorname{ABp}(\boldsymbol{\phi}_0(\widetilde{\boldsymbol{\phi}}_i), \ldots, \boldsymbol{\phi}_{-(p-1)}(\widetilde{\boldsymbol{\phi}}_i))\right\| \\ &\qquad + hL\sum_{j=1}^{p-2} |b_j| \left\|\widetilde{\boldsymbol{\phi}}_{i-j} - \boldsymbol{\phi}_{-j}(\widetilde{\boldsymbol{\phi}}_i)\right\| \\ &\quad\le |b_{-1}|AL^2C^*h^{p+3} + LBC^*h^{p+2}, \end{aligned} \tag{31}$$

where the last step follows by (23) for the first term and the induction hypothesis for the second.

Hence, (29), (30), and (31) deliver

$$\|\widetilde{\boldsymbol{\phi}}_{i+1} - \boldsymbol{\phi}_1(\widetilde{\boldsymbol{\phi}}_i)\| \le |b_{-1}|AL^2C^*h^{p+3} + LBC^*h^{p+2} + C_{\mathrm{PCp}}h^{p+1} \le Ch^{p+1},$$

because

$$|b_{-1}|AL^2C^*h^2 + LBC^*h + C_{\mathrm{PCp}} \le C$$

by (28). Hence, the bound (26) in (II) holds true for i replaced by $i+1$.

Let us now consider the bound (27) in (II) for i replaced by $i+1$. That this estimate holds true can be proved from the induction hypothesis exactly in the same way as in the proof of case (ii), which concludes the proof. □

We are now ready to study positively invariant sets for the ODE (1), that are also positively invariant for (i) RKp, (ii) ABp initialized with RKp, and (iii) PCp initialized with RKp.

3 Positively Invariant Sets

A positively invariant set for system (1), i.e. a set $P \subset \mathbb{R}^n$ such that $\boldsymbol{\phi}(t, \mathbf{x}) \in P$ for all $t \geq 0$ whenever $\mathbf{x} \in P$, is in general not positively invariant for a numerical procedure to approximate its solution trajectories. This is, for example, shown in [26, Example 3.1] for the system $\dot{\theta} = 1$, $\dot{r} = -r(1-r^2)$, in polar coordinates and the positively invariant set $\{(r, \theta) \in [0,1] \times [0, 2\pi)\}$.

We now show in Theorem 2 and Corollary 4 that for general systems, sublevel-sets $S \subset \mathbb{R}^n$ of certain Lyapunov-like functions V are positively invariant for both the system (1) and numerical methods to approximate its solution trajectories, if the step size $h > 0$ is sufficiently small.

Theorem 2. *(Positively invariant sets) Consider the system (1), let $V \in C^1(\mathbb{R}^n; \mathbb{R})$ and let S be a compact connected component of the sublevel set*

$$V^{-1}((-\infty, m]) := \{\mathbf{x} \in \mathbb{R}^n \colon V(\mathbf{x}) \leq m\}, \quad m \in \mathbb{R}.$$

Further assume that $\nabla V(\mathbf{x}) \cdot \mathbf{f}(\mathbf{x}) < 0$ and that $\nabla V(\mathbf{x})$ points out of S for every $\mathbf{x} \in \partial S$; that is, for every $\mathbf{x} \in \partial S$ there exists a $h > 0$ such that

$$\big[\mathbf{x} + (0,h)\nabla V(\mathbf{x})\big] \cap S = \emptyset.$$

where

$$\mathbf{x} + (0,h)\nabla V(\mathbf{x}) := \{\mathbf{y} \in \mathbb{R}^n \colon \mathbf{y} = \mathbf{x} + t\nabla V(\mathbf{x}),\ t \in (0,h)\}.$$

Then S is positively invariant for (1).

Further, assume that $\mathbf{f} \in C^p(\mathbb{R}^n; \mathbb{R}^n)$ and that we have a numerical method of order p in the sense of Definition 1. Then there is an $h' > 0$ such that if the time-step h of the numerical method fulfills $0 < h \leq h'$, then $\widetilde{\boldsymbol{\phi}}_{i+1}(\boldsymbol{\xi}) \in S$, whenever $\widetilde{\boldsymbol{\phi}}_k(\boldsymbol{\xi}) \in S$ for $k = 0, 1, 2, \ldots, i$, $i \in \mathbb{N}_0$.

Proof. Define

$$\delta := -\frac{1}{2} \max_{\mathbf{x} \in \partial S} \nabla V(\mathbf{x}) \cdot \mathbf{f}(\mathbf{x}) > 0,$$

i.e. $\nabla V(\mathbf{x}) \cdot \mathbf{f}(\mathbf{x}) \leq -2\delta$ for all $\mathbf{x} \in \partial S$, and let $\epsilon > 0$ be such that

$$\nabla V(\mathbf{x}) \cdot \mathbf{f}(\mathbf{x}) \leq -\delta < 0 \quad \text{for all} \quad \mathbf{x} \in \mathcal{W} := \{\mathbf{x} \in \mathbb{R}^n \colon d(\mathbf{x}, \partial S) \leq 2\epsilon\}.$$

To see that S is positively invariant for (1), consider that if it is not, then some solution trajectory starting in S must intersect ∂S at some point $\mathbf{x}$ and then leave S, that is, there exists an $\mathbf{x} \in \partial S$ and an $\tau^* > 0$ such that $\boldsymbol{\phi}(\tau, \mathbf{x}) \in \mathcal{W} \setminus S$ for all $0 < \tau \leq \tau^*$. Then

$$\begin{aligned} m < V(\boldsymbol{\phi}(\tau^*, \mathbf{x})) &= V(\mathbf{x}) + \int_0^{\tau^*} \frac{d}{d\tau} V(\boldsymbol{\phi}(\tau, \mathbf{x})) d\tau \\ &= m + \int_0^{\tau^*} \nabla V(\boldsymbol{\phi}(\tau, \mathbf{x})) \cdot \frac{d}{d\tau} \boldsymbol{\phi}(\tau, \mathbf{x}) d\tau \\ &= m + \int_0^{\tau^*} \nabla V(\boldsymbol{\phi}(\tau, \mathbf{x})) \cdot \mathbf{f}(\boldsymbol{\phi}(\tau, \mathbf{x})) d\tau \leq m - \delta\tau^*, \end{aligned}$$

a contradiction.

In order to prove that S is also positively invariant for the numerical method for small enough time-steps, set

$$\begin{aligned}
\mathcal{V} &:= \{\mathbf{x} \in \mathbb{R}^n \colon d(\mathbf{x}, \partial S) \le \epsilon\} \subset \mathcal{W}, \\
F &:= \max\{\max_{\mathbf{x}\in S} \|\mathbf{f}(\mathbf{x})\|_2, 1\}, \quad \text{and} \\
h' &:= \min\{\epsilon/(2\max\{F, C\}), 1, \tau^*, h^*\}.
\end{aligned}$$

Here and later in the proof, $C, h^* > 0$ are the constants for the numerical method from Definition 1.

Then, for $\mathbf{x} \in S \setminus \mathcal{V}$ and $0 \le h \le h'$ we have

$$\|\boldsymbol{\phi}(h, \mathbf{x}) - \mathbf{x}\|_2 \le \int_0^h \|\mathbf{f}(\boldsymbol{\phi}(s, \mathbf{x}))\|_2 ds \le hF \le \frac{\epsilon F}{2\max\{F, C\}} \le \frac{\epsilon}{2}$$

and it follows that

$$d(\boldsymbol{\phi}(h, \mathbf{x}), S \setminus \mathcal{V}) \le \epsilon/2, \quad \forall \mathbf{x} \in S \setminus \mathcal{V}. \tag{32}$$

Note that from (32), Definition 1 and for $\widetilde{\boldsymbol{\phi}}_k(\boldsymbol{\xi}) \in S$ for $k = 0, 1, 2, \ldots, i$, we have

$$\begin{aligned}
d(\widetilde{\boldsymbol{\phi}}_{i+1}(\boldsymbol{\xi}), S \setminus \mathcal{V}) &\le d(\boldsymbol{\phi}(h, \widetilde{\boldsymbol{\phi}}_i(\boldsymbol{\xi})), S \setminus \mathcal{V}) + \|\widetilde{\boldsymbol{\phi}}_{i+1}(\boldsymbol{\xi}) - \boldsymbol{\phi}(h, \widetilde{\boldsymbol{\phi}}_i(\boldsymbol{\xi}))\|_2 \\
&\le \epsilon/2 + \epsilon/2 = \epsilon,
\end{aligned}$$

using that $Ch^{p+1} \le \frac{\epsilon}{2}$ due to the definition of h'. Hence, $\widetilde{\boldsymbol{\phi}}_{i+1}(\boldsymbol{\xi}) \in S$ if the time-step of the numerical method fulfills $0 \le h \le h'$, whenever $\widetilde{\boldsymbol{\phi}}_k(\boldsymbol{\xi}) \in S$ for $k = 0, 1, 2, \ldots, i$ and additionally $\widetilde{\boldsymbol{\phi}}_i(\boldsymbol{\xi}) \in S \setminus \mathcal{V}$.

To finish the proof, we need to show the statement in the case $\widetilde{\boldsymbol{\phi}}_k(\boldsymbol{\xi}) \in S$ for $k = 0, 1, 2, \ldots, i$ and

$$\mathbf{x} = \widetilde{\boldsymbol{\phi}}_i(\boldsymbol{\xi}) \in S \cap \mathcal{V}.$$

We assume on the contrary that there are sequences $\boldsymbol{\xi}_j \in S$ and $0 < h_j \le h'$, $h_j \to 0$ as $j \to \infty$, such that

$$\widetilde{\boldsymbol{\phi}}^j_{i_j+1}(\boldsymbol{\xi}_j) \notin S$$

for all j, although

$$\widetilde{\boldsymbol{\phi}}^j_0(\boldsymbol{\xi}_j),\ \widetilde{\boldsymbol{\phi}}^j_1(\boldsymbol{\xi}_j),\ \ldots\ ,\widetilde{\boldsymbol{\phi}}^j_{i_j}(\boldsymbol{\xi}_j) \in S \quad \text{and} \quad \widetilde{\boldsymbol{\phi}}^j_{i_j}(\boldsymbol{\xi}_j) \in S \cap \mathcal{V}.$$

Here

$$\widetilde{\boldsymbol{\phi}}^j_0(\boldsymbol{\xi}_j),\ \widetilde{\boldsymbol{\phi}}^j_1(\boldsymbol{\xi}_j),\ \ldots\ ,\widetilde{\boldsymbol{\phi}}^j_{i_j}(\boldsymbol{\xi}_j),\ \widetilde{\boldsymbol{\phi}}^j_{i_j+1}(\boldsymbol{\xi}_j)$$

is the sequence generated by the numerical method with initial value $\boldsymbol{\xi}_j \in S$ and step-size h_j.

Note that since S is positively invariant for (1) we have $\boldsymbol{\phi}(h_j, \widetilde{\boldsymbol{\phi}}^j_{i_j}(\boldsymbol{\xi}_j)) \in S$ and therefore

$$V(\widetilde{\boldsymbol{\phi}}^j_{i_j}(\boldsymbol{\xi}_j)) \le m \quad \text{and} \quad V(\boldsymbol{\phi}(h_j, \widetilde{\boldsymbol{\phi}}^j_{i_j}(\boldsymbol{\xi}_j))) \le m \quad \text{for all } j.$$

Further, there exists $I \in \mathbb{N}$ such that for all $j \ge I$ we have

$$\boldsymbol{\phi}(\theta h_j, \widetilde{\boldsymbol{\phi}}^j_{i_j}(\boldsymbol{\xi}_j)) \in \mathcal{W} \cap S \quad \text{for all} \quad \theta \in [0,1] \quad \text{and} \quad V(\widetilde{\boldsymbol{\phi}}^j_{i_j+1}(\boldsymbol{\xi}_j)) > m.$$

Moreover, there is a convex and compact set $\widetilde{S} \supset S$ such that $\widetilde{\boldsymbol{\phi}}^j_{i_j+1}(\boldsymbol{\xi}_j) \in \widetilde{S}$ for all j. Let L_V be a Lipschitz constant for V on $\widetilde{S}$ and recall that by Definition 1 we have

$$\|\widetilde{\boldsymbol{\phi}}^j_{i_j+1}(\boldsymbol{\xi}_j) - \boldsymbol{\phi}(h_j, \widetilde{\boldsymbol{\phi}}^j_{i_j}(\boldsymbol{\xi}_j))\|_2 \le C h_j^{p+1}.$$

Now

$$\begin{aligned} \left| \frac{V(\widetilde{\boldsymbol{\phi}}^j_{i_j+1}(\boldsymbol{\xi}_j)) - V(\boldsymbol{\phi}(h_j, \widetilde{\boldsymbol{\phi}}^j_{i_j}(\boldsymbol{\xi}_j)))}{h_j} \right| &\le \frac{L_V \|\widetilde{\boldsymbol{\phi}}^j_{i_j+1}(\boldsymbol{\xi}_j) - \boldsymbol{\phi}(h_j, \widetilde{\boldsymbol{\phi}}^j_{i_j}(\boldsymbol{\xi}_j))\|}{h_j} \\ &\le \frac{L_V C h_j^{p+1}}{h_j} \\ &= L_V C h_j^p, \end{aligned} \tag{33}$$

$$\frac{V(\widetilde{\boldsymbol{\phi}}^j_{i_j+1}(\boldsymbol{\xi}_j)) - V(\widetilde{\boldsymbol{\phi}}^j_{i_j}(\boldsymbol{\xi}_j))}{h_j} > \frac{m - m}{h_j} = 0, \tag{34}$$

for all $j \ge I$.

Define $g_j(t) := V(\boldsymbol{\phi}(t, \widetilde{\boldsymbol{\phi}}^j_{i_j}(\boldsymbol{\xi}_j)))$. By the Mean-Value theorem there exists $\theta_j \in (0,1)$ such that

$$g_j(h_j) - g_j(0) = g'(\theta_j h_j) h_j \le -\delta h_j \tag{35}$$

holds for all $j \ge I$. From (34), (33), and (35) it follows that

$$\begin{aligned} 0 &\le \limsup_{j\to\infty} \frac{V(\widetilde{\boldsymbol{\phi}}^j_{i_j+1}(\boldsymbol{\xi}_j)) - V(\widetilde{\boldsymbol{\phi}}^j_{i_j}(\boldsymbol{\xi}_j))}{h_j} \\ &\le \limsup_{j\to\infty} \frac{V(\widetilde{\boldsymbol{\phi}}^j_{i_j+1}(\boldsymbol{\xi}_j)) - V(\boldsymbol{\phi}(h_j, \widetilde{\boldsymbol{\phi}}^j_{i_j}(\boldsymbol{\xi}_j)))}{h_j} \\ &\quad + \limsup_{j\to\infty} \frac{V(\boldsymbol{\phi}(h_j, \widetilde{\boldsymbol{\phi}}^j_{i_j}(\boldsymbol{\xi}_j))) - V(\widetilde{\boldsymbol{\phi}}^j_{i_j}(\boldsymbol{\xi}_j))}{h_j} \\ &= 0 + \limsup_{j\to\infty} \frac{g_j(h_j) - g_j(0)}{h_j} \\ &\le 0 - \delta < 0, \end{aligned}$$

which is a contradiction and thus the theorem is proved.

Theorem 2 establishes the existence of a positively invariant set for both the ODE and the numerical method to approximate its solution trajectories. To obtain such a set computationally, one can employ the RBF-CPA Lyapunov-like function. For this method one first computes Lyapunov-like function V by approximately solving Zubov's PDE

$$\nabla V(\mathbf{x}) \cdot \mathbf{f}(\mathbf{x}) = -g(\mathbf{x})$$

with a suitable positive function $g(\mathbf{x})$ using generalized interpolation in reproducing kernel Hilbert spaces and then interpolate its values over the simplices of a triangulation $\mathcal{T}$, see [19,24] for details.

In the following theorem, CPA$[\mathcal{T}]$ denotes the set of such interpolating functions, called continuous piecewise affine (CPA) functions, which are continuous and affine on each simplex of the triangulation $\mathcal{T}$. Further, $\nabla V_\nu \in \mathbb{R}^n$ denotes the constant gradient of V in the interior of a simplex $\mathfrak{S}_\nu \in \mathcal{T}$.

Theorem 3. *(CPA version of Thm. 2) Consider the system (1). Let $V \in \mathrm{CPA}[\mathcal{T}]$ and assume S is a compact connected component of $V^{-1}((-\infty, m])$, $m \in \mathbb{R}$. Assume that ∇V_ν points out of S at every $\mathbf{x} \in \partial S \cap \mathfrak{S}_\nu$ and that this is true for every $\mathfrak{S}_\nu \in \mathcal{T}$. Additionally, assume that there is a constant $c > 0$ such that $\nabla V_\nu \cdot \mathbf{f}(\mathbf{x}) \leq -c$ for every $\mathbf{x}$ in a neighbourhood of ∂S and every ν such that $\mathbf{x} \in \mathfrak{S}_\nu$, i.e. if $\mathbf{x} \in \mathfrak{S}_\nu \cap \mathfrak{S}_\mu$, then both $\nabla V_\nu \cdot \mathbf{f}(\mathbf{x}) \leq -c$ and $\nabla V_\mu \cdot \mathbf{f}(\mathbf{x}) \leq -c$. Then S is positively invariant for (1).*

Further, assume that $\mathbf{f} \in C^p(\mathbb{R}^n; \mathbb{R}^n)$ and that we have a numerical method of order p in the sense of Definition 1. Then there is an $h' > 0$ such that if the time-step h of the numerical method fulfills $0 < h \leq h'$, then $\widetilde{\boldsymbol{\phi}}_{i+1}(\boldsymbol{\xi}) \in S$, whenever $\widetilde{\boldsymbol{\phi}}_k(\boldsymbol{\xi}) \in S$ for $k = 0, 1, 2, \ldots, i$.

Proof. Essentially, the proof is the same as the proof of Theorem 2; the existence of $\delta = c/2$ and $\epsilon > 0$ now follow directly from the assumptions. The only reasoning that needs modification is why

$$g_j(h_j) - g_j(0) \leq -\delta h_j.$$

For $V \in \mathrm{CPA}[\mathcal{T}]$ this follows because for

$$D^+V(\mathbf{x}) := \limsup_{h \to 0+} \frac{V(\boldsymbol{\phi}(h, \mathbf{x})) - V(\mathbf{x})}{h}$$

we have

$$D^+V(\mathbf{x}) = \nabla V_\nu \cdot \mathbf{f}(\mathbf{x}) \leq -\delta,$$

where $\mathfrak{S}_\nu$ is a simplex such that $\mathbf{x} \in \mathfrak{S}_\nu$, see e.g. [32, Lem. 2.2], and by a generalized Mean Value Theorem, see [56] or [61, Thm. 12.24].

The following theorem uses the existence of a Lyapunov function to show the existence of a positively invariant set for both the flow and a numerical method satisfying Definition 1 by using Theorem 2.

Theorem 4. *(Positively invariant sets for the system and the numerical method) Let* $\mathbf{x}_0$ *be an exponentially stable equilibrium of* (1)*, where* $\mathbf{f} \in C^p(\mathbb{R}^n; \mathbb{R}^n)$ *with* $p \in \mathbb{N}$*, and let* $K \subset \mathcal{A}(\mathbf{x}_0)$ *be compact. Then there exists a compact and connected set* S*,* $K \subset S \subset \mathcal{A}(\mathbf{x}_0)$*, with the following property:*

Assume we have a numerical method of order p *in the sense of Definition 1. Then there exists a constant* $h' > 0$*, such that* S *is positively invariant both for the original flow* $\boldsymbol{\phi}(0, \boldsymbol{\xi}) = \boldsymbol{\xi} \in S$*,* $t \mapsto \boldsymbol{\phi}(t, \boldsymbol{\xi})$*, induced by* (1)*, and for the sequences* $\widetilde{\boldsymbol{\phi}}_i(\boldsymbol{\xi})$*,* $i \in \mathbb{N}_0$*, generated by the numerical method with step-size* h*,* $0 < h \leq h'$*, for the initial-values* $\boldsymbol{\xi} \in S$*. In other words,* $\widetilde{\boldsymbol{\phi}}_i(\boldsymbol{\xi}) \in S$ *for all* $\boldsymbol{\xi} \in S$ *and all* $i \in \mathbb{N}_0$*.*

Proof. By [14, Thm. 2.46] there exists a Lyapunov function V for the system (1) fulfilling $V(\mathbf{x}_0) = 0$ and

$$\nabla V(\mathbf{x}) \cdot \mathbf{f}(\mathbf{x}) = -\|\mathbf{x} - \mathbf{x}_0\|_2^2 \sqrt{1 + \|\mathbf{f}(\mathbf{x})\|_2^2}$$

for all $x \in \mathcal{A}(\mathbf{x}_0)$. We set $r := \max_{\mathbf{x} \in K} V(\mathbf{x})$ and $S := V^{-1}([0, r])$; hence $K \subset S$. Since V is also a Lyapunov function for the system

$$\dot{\mathbf{x}} = \mathbf{f}(\mathbf{x})(1 + \|\mathbf{f}(\mathbf{x})\|_2^2)^{-1/2},$$

with a bounded right-hand-side, the set $S \subset \mathcal{A}(\mathbf{x}_0)$ is compact. Since $V(\boldsymbol{\phi}(t, \boldsymbol{\xi})) \leq r$ for all $t \geq 0$ and

$$\mathbf{x}_0 \in \overline{\boldsymbol{\phi}([0, \infty), \boldsymbol{\xi})} \subset V^{-1}([0, r]) = S$$

for all $\boldsymbol{\xi} \in S$, the set S is also connected. Using this Lyapunov function and Theorem 2 for the numerical method, the existence of $h' > 0$ with the claimed properties follows.

Remark 2. By Theorem 1 any (i) Runge-Kutta method of order p (RKp), the (ii) Adams-Bashforth method of order p (ABp) initialized by RKp, and (iii) predictor-corrector methods of order p using ABp for the predictor step and the Adams-Moulton method of order p for the corrector step, $p \in \mathbb{N}$, are all numerical methods of order p in the sense of Definition 1, and thus, fulfill the assumptions in Theorem 4. These results are used in [21] and [25] to prove that Lyapunov functions and contraction metrics can be approximated arbitrarily close on compact sets using numerical integration and quadrature.

4 Example

We demonstrate our results by computing a positively invariant set and a contraction metric for an SIRS model, which includes susceptible (S), infected (I), and recovered (R) individuals. It is an extension of the well known SIR model, but where recovered individuals can lose their immunity after a certain period and become susceptible again. The SIRS model is particularly relevant in the context of diseases for which vaccination is available and plays a significant role in disease dynamics. It is described by the following equations:

$$\begin{cases} \dot{S} = -\beta \frac{SI}{N} + \mu R \\ \dot{I} = \beta \frac{SI}{N} - \gamma I \\ \dot{R} = \gamma I - \mu R \end{cases}$$

where

- S is the number of susceptible individuals,
- I is the number of infected individuals,
- R is the number of recovered (and temporarily immune) individuals,
- β is the transmission rate of the disease,
- γ is the recovery rate,
- μ is the rate at which recovered individuals lose immunity and return to the susceptible state.
- N is the total population (assumed constant, $N = S + I + R$).

The SIRS model's equilibria will depend on the parameter values and can include disease-free equilibria (where $I = 0$) and endemic equilibria (where the disease persists in the population). We set the parameters as:

- $\beta = 0.3$
- $\gamma = 0.1$
- $\mu = 0.02$

The dimension of the model can be decreased by one by replacing R with $N - S - I$ (since the total population is assumed to be constant). Further, we use the proportional variables $x := \frac{S}{N}$ and $y := \frac{I}{N}$ to improve the visibility of the solutions. This results in the following equivalent 2-dimensional system:

$$\begin{cases} \dot{x} = -\beta xy + \mu(1 - x - y) \\ \dot{y} = \beta xy - \gamma y \end{cases} \tag{36}$$

The system has a disease-free repelling equilibrium at $(1, 0)$, and an endemic asymptotically stable equilibrium at $\mathbf{x}_0 := \left(\frac{\gamma}{\beta}, \frac{\mu}{\beta}\frac{\beta-\gamma}{\gamma+\mu}\right) \approx (0.33, 0.11)$. In order to compute a positively invariant set S for this dynamical system as in Theorem 3 we used the method described in [24] and motivated by [19]. In this method one first solves numerically the Zubov-like PDE $\nabla V(\mathbf{x}) \cdot \mathbf{f}(\mathbf{x}) = -\|\mathbf{f}(\mathbf{x}) - \mathbf{x}_0\|_2$ using collocation with radial basis functions, where $\mathbf{x}_0$ is the (asymptotically stable) equilibrium under consideration and $\mathbf{f}$ is the right-hand side function in (36). We used $9,835$ collocation points with a hexagonal grid to cover the area $[-0.667, 1.333] \times [-0.769, 0.991]$ except a circle of radius 0.2 around $(1, 0)$. Then we interpolate the numerical solution V by a CPA function V_P on a simplicial complex to obtain a function with negative orbital derivative, except in the area plotted in yellow in Figure 1. Sublevel-sets of the function V_P can now be used to obtain sets, that are positively invariant, both for the ODE (36) and for numerical methods to approximate its solutions. Such a positively invariant set S is plotted in cyan in Figure 1 around the equilibrium $\mathbf{x}_0$ (the green points).

The results of this paper were used to prove that we can always compute a contraction metric using numerical integration, see [25]. As an example, we use here the AB4 multi-step method (initialized with RK4) for numerical integration as well as quadrature of the solution trajectories of (36) to compute a contraction metric. The conditions for the contraction metric are fulfilled in the white area, or equivalently, the area where it fails the conditions is depicted in red and blue. Since the sublevel set of V_P has empty intersection with the red and blue area, the sublevel set is a subset of the basin of attraction of the equilibrium.

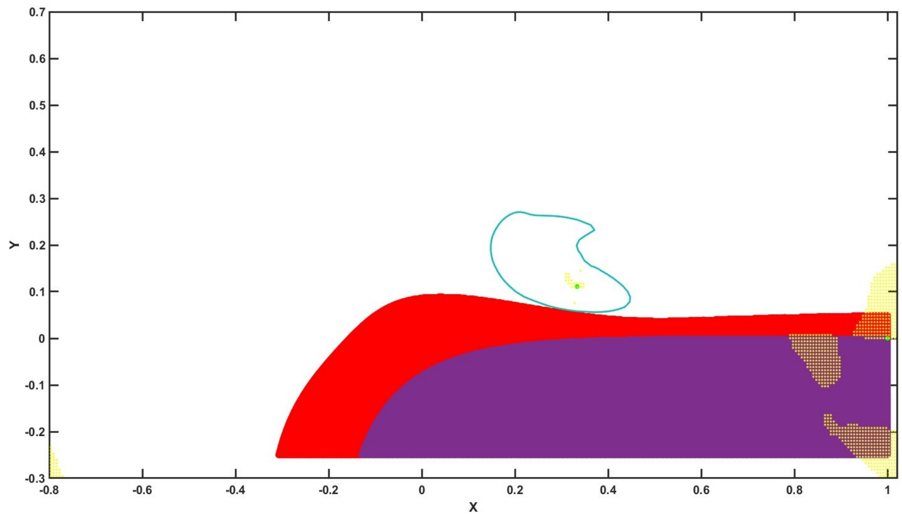

Fig. 1: The system (36). The green point at $(0.33, 0.11)$ is the asymptotically stable equilibrium $\mathbf{x}_0$ and a positively invariant sublevel set of V_P is drawn in cyan around it. The yellow area is where the orbital derivative of V_P is not negative, the red area is where the computed contraction metric fails to have a negative definite derivative, and the blue area is where it fails to be positive definite. In the rest of the area it is a contraction metric. This shows that the sublevel set inside the cyan curve is a subset of the basin of attraction of $\mathbf{x}_0$.

5 Conclusions

We have shown that for an ODE with an exponentially stable equilibrium $\mathbf{x}_0$ and any compact subset K of its basin of attraction $\mathcal{A}(\mathbf{x}_0)$, we can find a compact and connected set S, $K \subset S \subset \mathcal{A}(\mathbf{x}_0)$, that is positively invariant, both for the ODE and its numerical approximation. We have established that this property holds for a variety of established numerical methods, namely Runge-Kutta, Adams-Bashforth, and predictor-corrector methods based on Adams-Bashforth and Adams-Moulton, all of arbitrary order. Finally, we demonstrated a constructive way of how to compute such positively invariant sets in practice and have shown this in an example.

Acknowledgements This work was supported in part by the Icelandic Research Fund under Grant 228725-051.

References

1. Aminzare, Z., Sontag, E.: Contraction methods for nonlinear systems: A brief introduction and some open problems. In: Proceedings of the 53rd IEEE Conference on Decision and Control. pp. 3835–3847 (2014). https://doi.org/10.1109/CDC.2014.7039986
2. Anderson, J., Papachristodoulou, A.: Advances in computational Lyapunov analysis using sum-of-squares programming. Discrete Contin. Dyn. Syst. Ser. B **20**(8), 2361–2381 (2015)
3. Aylward, E., Parrilo, P., Slotine, J.J.: Stability and robustness analysis of nonlinear systems via contraction metrics and SOS programming. Automatica **44**(8), 2163–2170 (2008). https://doi.org/10.1016/j.automatica.2007.12.012, https://doi.org/10.1016/j.automatica.2007.12.012
4. Björnsson, J., Giesl, P., Hafstein, S., Kellett, C., Li, H.: Computation of continuous and piecewise affine Lyapunov functions by numerical approximations of the Massera construction. In: Proceedings of the CDC, 53rd IEEE Conference on Decision and Control. pp. 5506–5511. Los Angeles (CA), USA (2014)
5. Björnsson, J., Giesl, P., Hafstein, S., Kellett, C., Li, H.: Computation of Lyapunov functions for systems with multiple attractors. Discrete Contin. Dyn. Syst. Ser. A **35**(9), 4019–4039 (2015)
6. Björnsson, J., Hafstein, S.: Efficient Lyapunov function computation for systems with multiple exponentially stable equilibria. Procedia Computer Science **108**, 655–664 (2017), proceedings of the International Conference on Computational Science (ICCS), Zurich, Switzerland, 2017.
7. Borg, G.: A condition for the existence of orbitally stable solutions of dynamical systems. Kungl. Tekn. Högsk. Handl. 153 (1960)
8. Bullo, F.: Contraction Theory for Dynamical Systems. Kindle Direct Publishing, 1 edn. (2023)
9. Demidovič, B.: On the dissipativity of a certain non-linear system of differential equations. I. Vestnik Moskov. Univ. Ser. I Mat. Meh. **1961**(6), 19–27 (1961)
10. Deuflhard, P., Hohmann, A.: Numerische Mathematik 2. de Gruyter, 4th edn. (2008)
11. Doban, A.: Stability domains computation and stabilization of nonlinear systems: implications for biological systems. PhD thesis: Eindhoven University of Technology (2016)
12. Doban, A., Lazar, M.: Computation of Lyapunov functions for nonlinear differential equations via a Yoshizawa-type construction. IFAC-PapersOnLine **49**(18), 29 – 34 (2016)
13. Forni, F., Sepulchre, R.: A differential Lyapunov framework for Contraction Analysis. IEEE Transactions on Automatic Control **59**(3), 614–628 (2014)
14. Giesl, P.: Construction of Global Lyapunov Functions Using Radial Basis Functions. Lecture Notes in Math. 1904, Springer (2007)

15. Giesl, P.: Converse theorems on contraction metrics for an equilibrium. J. Math. Anal. Appl. (2015), in press, doi:10.1016/j.jmaa.2014.12.010
16. Giesl, P.: Computation of a contraction metric for a periodic orbit using meshfree collocation. SIAM J. Appl. Dyn. Syst. **18**(3), 1536–1564 (2019)
17. Giesl, P., Hafstein, S.: Construction of a CPA contraction metric for periodic orbits using semidefinite optimization. Nonlinear Anal. **86**, 114–134 (2013)
18. Giesl, P., Hafstein, S.: Revised CPA method to compute Lyapunov functions for nonlinear systems. J. Math. Anal. Appl. **410**(1), 292–306 (2014)
19. Giesl, P., Hafstein, S.: Computation and verification of Lyapunov functions. SIAM J. Appl. Dyn. Syst. **14**(4), 1663–1698 (2015)
20. Giesl, P., Hafstein, S.: Review of computational methods for Lyapunov functions. Discrete Contin. Dyn. Syst. Ser. B **20**(8), 2291–2331 (2015)
21. Giesl, P., Hafstein, S.: Dynamical Systems in Theoretical Perspective. In: Awrejcewicz, J. (ed.) Springer Proceedings in Mathematics and Statistics, chap. Lyapunov Functions by Interpolating Numerical Quadrature: Proof of Convergence, p. (to appear). Springer (2023), springer Proceedings in Mathematics and Statistics
22. Giesl, P., Hafstein, S., Haraldsdottir, M., Thorsteinsson, D., Kawan, C.: Subgradient algorithm for computing contraction metrics for equilibria. J. Comput. Dynamics **10**(2), 281–303 (2023)
23. Giesl, P., Hafstein, S., Kawan, C.: Review on contraction analysis and computation of contraction metrics. J. Comput. Dynamics **10**(1), 1–47 (2023)
24. Giesl, P., Hafstein, S., Mehrabinezhad, I.: Computation and verification of contraction metrics for periodic orbits. J. Math. Anal. Appl. **503**(2), Paper No. 125309, 32 (2021)
25. Giesl, P., Hafstein, S., Mehrabinezhad, I.: Contraction metrics by numerical integration and quadrature: Uniform error estimate. In: Proceedings of the 20th International Conference on Informatics in Control, Automation and Robotics (ICINCO). p. (submitted) (2023)
26. Giesl, P., Hafstein, S., Mehrabinezhad, I.: Positively invariant sets for ODEs and numerical integration. In: Proceedings of the 20th International Conference on Informatics in Control, Automation and Robotics (ICINCO). p. (submitted) (2023)
27. Giesl, P., Hafstein, S., Mehrabinezhad, I.: Contraction metric computation using numerical integration and quadrature. Discrete Contin. Dyn. Syst. Ser. B **29**(6), 2610–2632 (2024). https://doi.org/10.1016/j.jmaa.2021.125309, https://doi.org/10.1016/j.jmaa.2021.125309
28. Gudmundsson, S., Hafstein, S.: Lyapunov function verification: MATLAB implementation. In: Proceedings of the 1st Conference on Modelling, Identification and Control of Nonlinear Systems (MICNON). pp. 806–811. No. 0235 (2015)
29. Hafstein, S.: A constructive converse Lyapunov theorem on exponential stability. Discrete Contin. Dyn. Syst. Ser. A **10**(3), 657–678 (2004)
30. Hafstein, S.: A constructive converse Lyapunov theorem on asymptotic stability for nonlinear autonomous ordinary differential equations. Dynamical Systems: An International Journal **20**(3), 281–299 (2005)
31. Hafstein, S.: Numerical analysis project in ODEs for undergraduate students. In: Rodrigues, J., Cardoso, P., Monteiro, J., Lam, R., Krzhizhanovskaya, V., Lees, M., Dongarra, J., Sloot, P. (eds.) Computational Science - ICCS 2019: 19th International Conference, Faro, Portugal, June 12-14, 2019, Proceedings, Part V, pp. 412–434. Springer (2019)
32. Hafstein, S.: CPA Lyapunov functions: Switched systems vs. differential inclusions. In: Proceedings of the 17th International Conference on Informatics in Control, Automation and Robotics (ICINCO). pp. 745–753 (2020)

33. Hafstein, S., Kellett, C., Li, H.: Computation of Lyapunov functions for discrete-time systems using the Yoshizawa construction. In: Proceedings of 53rd IEEE Conference on Decision and Control (CDC) (2014)
34. Hafstein, S., Kellett, C., Li, H.: Continuous and piecewise affine Lyapunov functions using the Yoshizawa construction. In: Proceedings of the 2014 American Control Conference (ACC). pp. 548–553 (no. 0170). Portland (OR), USA (2014)
35. Hafstein, S., Kellett, C., Li, H.: Computing continuous and piecewise affine Lyapunov functions for nonlinear systems. J. Comput. Dyn **2**(2), 227 – 246 (2015)
36. Hafstein, S., Valfells, A.: Study of dynamical systems by fast numerical computation of Lyapunov functions. In: Proceedings of the 14th International Conference on Dynamical Systems: Theory and Applications (DSTA). vol. Mathematical and Numerical Aspects of Dynamical System Analysis, pp. 220–240 (2017)
37. Hafstein, S., Valfells, A.: Efficient computation of Lyapunov functions for nonlinear systems by integrating numerical solutions. Nonlinear Dynamics **97**(3), 1895–1910 (2019)
38. Hahn, W.: Stability of Motion. Springer, Berlin (1967)
39. Hairer, E., Nørsett, S., Wanner, G.: Solving Ordinary Differential Equations I: Nonstiff Problems. Springer, 3rd edn. (2008)
40. Hartman, P.: On stability in the large for systems of ordinary differential equations. Canadian J. Math. **13**, 480–492 (1961). https://doi.org/10.4153/CJM-1961-040-6, https://doi.org/10.4153/CJM-1961-040-6
41. Hartman, P.: Ordinary Differential Equations. Wiley, New York (1964)
42. Julian, P., Guivant, J., Desages, A.: A parametrization of piecewise linear Lyapunov functions via linear programming. Int. J. Control **72**(7-8), 702–715 (1999)
43. Kamyar, R., Peet, M.: Polynomial optimization with applications to stability analysis and control – an alternative to sum of squares. Discrete Contin. Dyn. Syst. Ser. B **20**(8), 2383–2417 (2015)
44. Khalil, H.: Nonlinear Systems. Pearson, 3. edn. (2002)
45. Krasovskiĭ, N.N.: Problems of the Theory of Stability of Motion. Mir, Moskow, 1959. English translation by Stanford University Press (1963)
46. Lewis, D.: Differential equations referred to a variable metric. Amer. J. Math. **73**, 48–58 (1951)
47. Lewis, D.C.: Metric properties of differential equations. Amer. J. Math. **71**, 294–312 (1949), http://www.jstor.org/stable/2372245
48. Li, H., Hafstein, S., Kellett, C.: Computation of continuous and piecewise affine Lyapunov functions for discrete-time systems. J. Difference Equ. Appl. **21**(6), 486–511 (2015)
49. Lohmiller, W., Slotine, J.J.: On Contraction Analysis for Non-linear Systems. Automatica **34**, 683–696 (1998)
50. Marinósson, S.: Lyapunov function construction for ordinary differential equations with linear programming. Dynamical Systems: An International Journal **17**(2), 137–150 (2002)
51. Parrilo, P.: Structured Semidefinite Programs and Semialgebraic Geometry Methods in Robustness and Optimization. PhD thesis: California Institute of Technology Pasadena, California (2000)
52. Piccini, J., August, E., Hafstein, S., Andersen, A.: Sufficient stability conditions for a class of switched systems with multiple steady states. IEEE Contr. Syst. Lett. **7**, 2653–2658 (2023)
53. Ratschan, S., She, Z.: Providing a basin of attraction to a target region of polynomial systems by computation of Lyapunov-like functions. SIAM J. Control Optim. **48**(7), 4377–4394 (2010)

54. Sastry, S.: Nonlinear Systems: Analysis, Stability, and Control. Springer (1999)
55. Sauer, T.: Numerical Analysis. Pearson, 2nd edn. (2012)
56. Scheeffer, L.: Zur Theorie der stetigen Funktionen einer reellen Veränderlichen. Acta Math. **5**(1), 183–194, 279–296 (1884)
57. Simpson-Porco, J., Bullo, F.: Contraction theory on Riemannian manifolds. Systems Control Lett. **65**, 74–80 (2014). https://doi.org/10.1016/j.sysconle.2013.12.016, https://doi.org/10.1016/j.sysconle.2013.12.016
58. Valmorbida, G., Anderson, J.: Region of attraction estimation using invariant sets and rational Lyapunov functions. Automatica **75**, 37–45 (2017)
59. Vannelli, A., Vidyasagar, M.: Maximal Lyapunov functions and domains of attraction for autonomous nonlinear systems. Automatica **21**(1), 69–80 (1985)
60. Vidyasagar, M.: Nonlinear System Analysis. Classics in Applied Mathematics, SIAM, 2. edn. (2002)
61. Walter, W.: Analysis 1. Springer, 7th edn. (2004)
62. Yoshizawa, T.: Stability theory by Liapunov's second method. Publications of the Mathematical Society of Japan, No. 9, The Mathematical Society of Japan, Tokyo (1966)
63. Zubov, V.I.: Methods of A. M. Lyapunov and their application. Translation prepared under the auspices of the United States Atomic Energy Commission; edited by Leo F. Boron, P. Noordhoff Ltd, Groningen (1964)