# Dynamics of Adaptive Time-Stepping ODE solvers

A.R. Humphries

Tony.Humphries@mcgill.ca

McGill

Joint work with:

NICK CHRISTODOULOU

PAUL FENTON

RATNAM VIGNESWARAN

ARH acknowledges support of the Engineering and Physical Sciences Research Council (UK), Leverhulme Trust (UK), and McGill.

## Abstract

For efficiency, variable time-stepping methods are often used to numerically integrate dynamical systems. The flow on chaotic attractors is often organised by the unstable manifolds of the fixed points, and it is thus necessary to obtain good numerical approximations in the neighbourhood of fixed points to reproduce the dynamics. However the standard adaptive algorithm typically fails to do this. Implicit methods designed for stiff problems are also unsuitable; they typically destroy the structure of the unstable manifold unless very small step-sizes are used. We will present examples to illustrate these poor dynamical behaviours, together with theoretical results on the approximation of stable/unstable manifolds, and suggest a phase space/stability based improvement to the standard algorithm.

# Contents

# Approximation of a dynamical system with a fixed step-size Runge-Kutta method

The dynamical system

$$\dot{u}(t) = f(u(t)), \qquad u(0) = U \in \mathbb{R}^d,$$

has solution operator $S(\bullet)$ and so $u(t) = S(t)U$ for all $t$.

# Approximation of a dynamical system with a fixed step-size Runge-Kutta method

The dynamical system

$$\dot{u}(t) = f(u(t)), \qquad u(0) = U \in \mathbb{R}^d,$$

has solution operator $S(\bullet)$ and so $u(t) = S(t)U$ for all $t$.

$$\begin{array}{c|c} c & \mathcal{A} \\ \hline & b^T \end{array}$$

The order $p$ Runge-Kutta method has evolution map $S_h$.

Example Forward Euler is defined by

$$u_{n+1} = u_n + hf(u_n) \equiv S_h(u_n), \qquad \forall n \geqslant 0, \qquad u_0 = U.$$

## Approximation of a dynamical system with a fixed step-size Runge-Kutta method

The dynamical system

$$\dot{u}(t) = f(u(t)), \qquad u(0) = U \in \mathbb{R}^d,$$

has solution operator $S(\bullet)$ and so $u(t) = S(t)U$ for all $t$.

$$\begin{array}{c|c} c & \mathcal{A} \\ \hline & b^T \end{array}$$

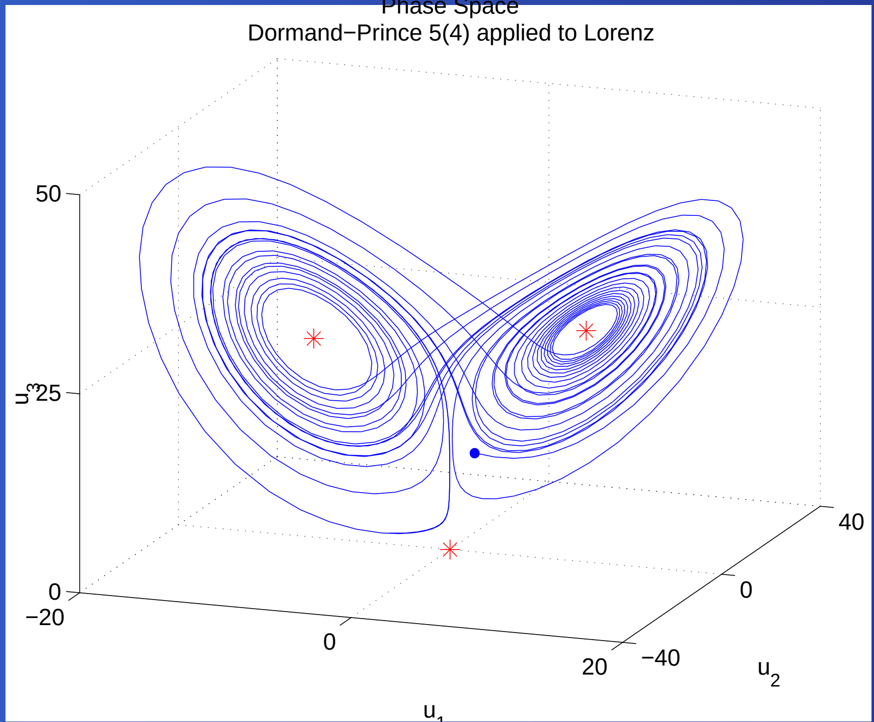
The order $p$ Runge-Kutta method has evolution map $S_h$.

$S_h(u)$ advances the numerical solution with step-size $h$

$$u_{n+1} = S_h(u_n), \qquad \forall n \geqslant 0, \qquad u_0 = U.$$

Each $u_n$ is an approximation of $S(nh)U = u(nh)$.

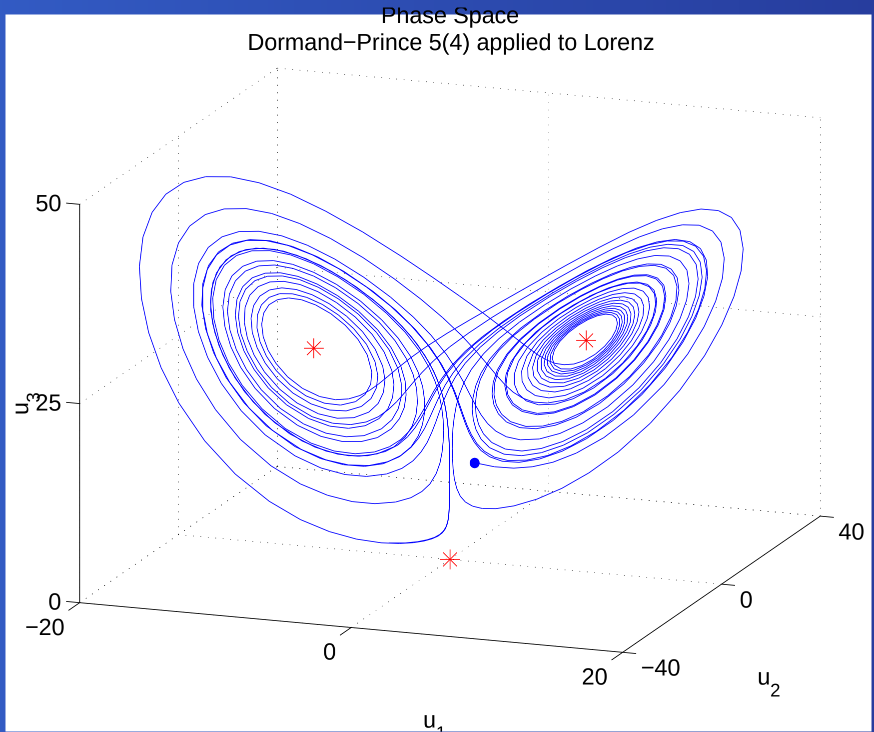
# Organisation of the flow




Flow in forward time organised by fixed points and unstable manifolds.

So consider approximation of unstable manifolds by numerical methods.

# Organisation of the flow




Flow in forward time organised by fixed points and unstable manifolds.

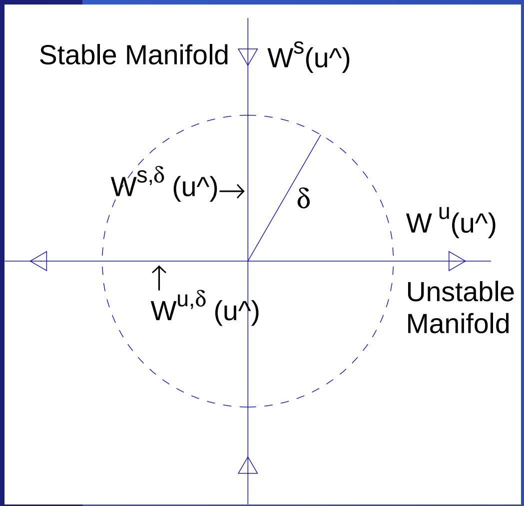
So consider approximation of unstable manifolds by numerical methods.

Lorenz vector field is

$$f(x, y, z) = (\sigma(y - x), \quad rx - y - xz, \quad xy - bz)\,.$$

Relationship of flow to fixed points obvious from figure.

# Approximation of Local Unstable Manifolds with a Fixed Step-Size Runge-Kutta Method



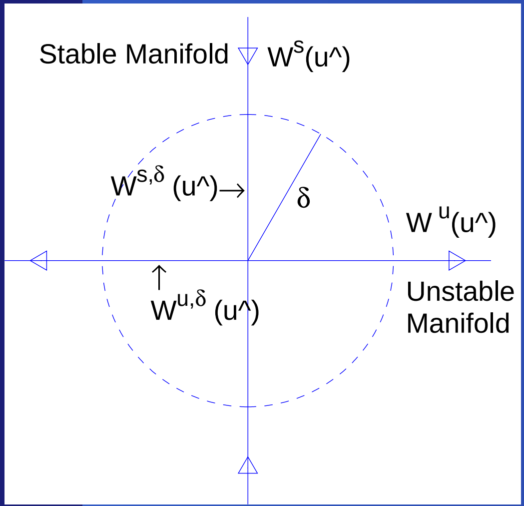

The unstable manifold of equilibrium point $\widehat{u}$ is

$$W^u(\widehat{u}) = \{u \in \mathbb{R}^d : \|S(-t)u - \widehat{u}\| \to 0 \text{ as } t \to \infty\}.$$

Let $\delta > 0$. The local unstable manifold of $\widehat{u}$ is

$$W^{u,\delta}(\widehat{u}) = \{u \in W^u(\widehat{u}) : \|S(-t)u - \widehat{u}\| \leqslant \delta \;\; \forall t \geqslant 0\}.$$

# Approximation of Local Unstable Manifolds with a Fixed Step-Size Runge-Kutta Method




The unstable manifold of equilibrium point $\widehat{u}$ is

$$W^u(\widehat{u}) = \{u \in \mathbb{R}^d : \|S(-t)u - \widehat{u}\| \to 0 \text{ as } t \to \infty\}.$$

Let $\delta > 0$. The local unstable manifold of $\widehat{u}$ is

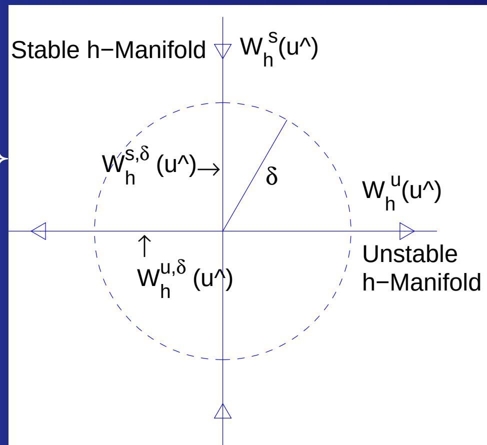
$$W^{u,\delta}(\widehat{u}) = \{u \in W^u(\widehat{u}) : \|S(-t)u - \widehat{u}\| \leqslant \delta \ \ \forall t \geqslant 0\}.$$

Generate $\{u_n\}_{n \geqslant 0}$ using a fixed step-size $h$. The unstable $h$-manifold of $\widehat{u}$ is

$$W_h^u(\widehat{u}) = \{u \in \mathbb{R}^d | u_0 = u, \|u_{-k} - \widehat{u}\| \to 0 \text{ as } k \to \infty\}$$

Let $\delta > 0$. The local unstable $h$-manifold of $\widehat{u}$ is

$$W_h^{u,\delta}(\widehat{u}) = \{u \in W_h^u(\widehat{u}) | u_0 = u, \|u_{-k} - \widehat{u}\| \leqslant \delta \ \forall k\}.$$




McGill

LaTeX

# Local Unstable Manifold Theorem

Theorem Apply a fixed step-size Runge-Kutta method of order $p$ to $\dot{u} = f(u)$. Let $\widehat{u}$ be a hyperbolic equilibrium and $f \in C^{p+1}(\mathbb{R}^d)$. Then there exists $C, H, \Delta > 0$ such that $\forall \delta \in (0, \Delta)$, $\forall h \in (0, H)$ the following holds:

for each $u \in W^{u,\delta}(\widehat{u})$, there exists a $u_h \in W_h^{u,\delta}(\widehat{u})$ such that

$$\|u - u_h\| \leqslant Ch^p \|u - \widehat{u}\|^2;$$

and for each $u_h \in W_h^{u,\delta}(\widehat{u})$, there exists $u \in W^{u,\delta}(\widehat{u})$ such that

$$\|u - u_h\| \leqslant Ch^p \|u_h - \widehat{u}\|^2.$$

## Proof

- Let $Df(\widehat{u})$ be the Jacobian of $f$ evaluated at $\widehat{u}$.
- Shift the coordinates $v = u - \widehat{u}$.
- Linearise the solution operator and the Runge-Kutta evolution map about 0

$$S(h)v = \exp(hDf(\widehat{u}))v + G_h(v), \quad \widehat{S}_h(v) = R(hDf(\widehat{u}))v + N_h(v),$$

  where $R$ is matrix generalisation of linear stability function.
- Show that both $W^{u,\delta}(\widehat{u})$ and $W_h^{u,\delta}(\widehat{u})$ are indeed manifolds.
- Show that both $W^{u,\delta}(\widehat{u})$ and $W_h^{u,\delta}(\widehat{u})$ are representable as graphs.
- Show that the graphs are close. $\square$

## Remarks

- A parallel result holds for local stable manifolds.

McGill

LaTeX

# Remarks

- A parallel result holds for local stable manifolds.
- The result is a generalisation of Beyn's result: there exists $C > 0$ such that
$$\|u - u_h\| \leqslant Ch^p.$$

## Remarks

- A parallel result holds for local stable manifolds.
- The result is a generalisation of Beyn's result: there exists $C > 0$ such that

$$\|u - u_h\| \leqslant Ch^p.$$

- Theorem 1 implies that $W^{u,\delta}(\widehat{u})$ and $W_h^{u,\delta}(\widehat{u})$ are tangential at the fixed point, and so is a form of local (un)stable manifold theorem. This follows from $\| \bullet - \widehat{u}\|^2$ on RHS of equations; so distance between numerical and exact manifolds depends on the square of the distance from the fixed point.

# Approximation of a Dynamical System with a Variable Step-Size Runge-Kutta Embedded Pair

Consider embedded Runge-Kutta pair with $|p - \widetilde{p}| = 1$.

$$\begin{array}{r|l} c & \mathcal{A} \\ \hline u_{n+1} = S_{h_n}(u_n) & b^T \\ \widetilde{u}_{n+1} = \widetilde{S}_{h_n}(u_n) & \widetilde{b}^T \end{array} \qquad \begin{array}{l} \\ \text{order } p \\ \text{order } \widetilde{p} \end{array}$$

# Approximation of a Dynamical System with a Variable Step-Size Runge-Kutta Embedded Pair

Consider embedded Runge-Kutta pair with $|p - \widetilde{p}| = 1$.

$$
\begin{array}{r|l}
c & \mathcal{A} \\
\hline
u_{n+1} = S_{h_n}(u_n) & b^T \\
\widetilde{u}_{n+1} = \widetilde{S}_{h_n}(u_n) & \widetilde{b}^T
\end{array}
\qquad
\begin{array}{l}
\\
\text{order } p \\
\text{order } \widetilde{p}
\end{array}
$$

$S_h(u)$ advances the numerical solution

$$u_{n+1} = S_{h_n}(u_n), \qquad \forall n \geqslant 0, \qquad u_0 = U.$$

# Approximation of a Dynamical System with a Variable Step-Size Runge-Kutta Embedded Pair

Consider embedded Runge-Kutta pair with $|p - \widetilde{p}| = 1$.

$$\begin{array}{r|l} c & \mathcal{A} \\ \hline u_{n+1} = S_{h_n}(u_n) & b^T \\ \widetilde{u}_{n+1} = \widetilde{S}_{h_n}(u_n) & \widetilde{b}^T \end{array} \qquad \begin{array}{l} \\ \text{order } p \\ \text{order } \widetilde{p} \end{array}$$

$S_h(u)$ advances the numerical solution

$$u_{n+1} = S_{h_n}(u_n), \qquad \forall n \geqslant 0, \qquad u_0 = U.$$

Note

Does not define dynamical system on $\mathbb{R}^d$ as $h_n$ varies with $n$.

# Local error approximation

With user-defined tolerance, $0 < \tau \ll 1$, step $h_n$ chosen by

$$\|E(u_n, h_n)\| \leqslant \tau, \quad \text{where} \quad E(u_n, h_n) = \frac{1}{h_n^\rho}(u_{n+1} - \widetilde{u}_{n+1}).$$

with $\rho = 0$ error per step (EPS) or $\rho = 1$ error per unit step (EPUS).

## Local error approximation

With user-defined tolerance, $0 < \tau \ll 1$, step $h_n$ chosen by

$$\|E(u_n, h_n)\| \leqslant \tau, \quad \text{where} \quad E(u_n, h_n) = \frac{1}{h_n^\rho}(u_{n+1} - \widetilde{u}_{n+1}).$$

with $\rho = 0$ error per step (EPS) or $\rho = 1$ error per unit step (EPUS).

Algorithm attempts to ensure

$$E(u_n, h_n) \approx \gamma\tau, \quad \gamma \in (0, 1) \text{ safety factor}$$

Leads to trouble near fixed points since $f(u_n) = 0$ implies

$$E(u_n, h_n) = 0.$$

McGill

LaTeX

# Step-changing Algorithm

Let $h_n^k$ be a candidate for $h_n$.

- If error control condition is satisfied, set $h_n = h_n^k$, update solution and set

$$h_{n+1}^0 = \min\left[h_{\max}, \left(\frac{\gamma\tau}{\|E(u_n, h_n)\|}\right)^{(1/\overline{p})} h_n\right].$$

- If error control condition not satisfied set

$$h_n^{k+1} = \left(\frac{\gamma\tau}{\|E(u_n, h_n^k)\|}\right)^{(1/\overline{p})} h_n^k.$$

$\overline{p} = \min(p, \widetilde{p}) + 1 - \rho$ $\qquad$ $\gamma \in (0, 1)$ safety factor.

# Adaptive Time-Stepping Algorithm as Dynamical System

Let $\tau > 0$ be a (user-defined) error tolerance.
An acceptable step-size $h$ for $u \in \mathbb{R}^d$ satisfies

$$\|S_h(u) - \widetilde{S}_h(u)\| \leqslant \tau.$$

The maximum step-size (independent of $\tau$) is $h_{\max}$.

# Adaptive Time-Stepping Algorithm as Dynamical System

Let $\tau > 0$ be a (user-defined) error tolerance.
An acceptable step-size $h$ for $u \in \mathbb{R}^d$ satisfies

$$\|S_h(u) - \widetilde{S}_h(u)\| \leqslant \tau.$$

The maximum step-size (independent of $\tau$) is $h_{\max}$.
Construct the sequence $\{(u_n, h_n)\}_{n \geqslant 0}$ using the algorithmic map

$$S_\tau : \mathbb{R}^d \times (0, h_{\max}] \longrightarrow \mathbb{R}^d \times (0, h_{\max}]$$

$$(u_{n+1}, h_{n+1}) = S_\tau(u_n, h_n).$$

$S_\tau$ finds an acceptable step-size $h_n$ and advances the solution.

# Variable Time-Stepping Algorithm as Dynamical System

$$S_\tau : \mathbb{R}^d \times (0, h_{\max}] \longrightarrow \mathbb{R}^d \times (0, h_{\max}]$$

$$(u_{n+1}, h_{n+1}) = S_\tau(u_n, h_n).$$

# Variable Time-Stepping Algorithm as Dynamical System

$$S_\tau : \mathbb{R}^d \times (0, h_{\max}] \longrightarrow \mathbb{R}^d \times (0, h_{\max}]$$

$$(u_{n+1}, h_{n+1}) = S_\tau(u_n, h_n).$$

- $S_\tau$ is discontinuous.

# Variable Time-Stepping Algorithm as Dynamical System

$$S_\tau : \mathbb{R}^d \times (0, h_{\max}] \longrightarrow \mathbb{R}^d \times (0, h_{\max}]$$

$$(u_{n+1}, h_{n+1}) = S_\tau(u_n, h_n).$$

- $S_\tau$ is discontinuous.
- (A. Stuart & H. Lamba) There exists $\gamma \in (0,1)$ and $C > 0$ such that

$$h \leqslant C \left( \frac{\tau}{\|f(u)\|} \right)^\gamma .$$

# Variable Time-Stepping Algorithm as Dynamical System

$$S_\tau : \mathbb{R}^d \times (0, h_{\max}] \longrightarrow \mathbb{R}^d \times (0, h_{\max}]$$

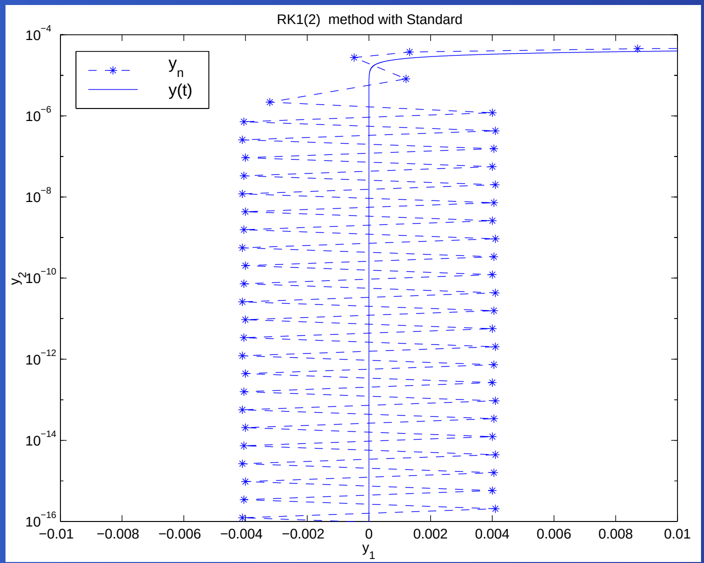
$$(u_{n+1}, h_{n+1}) = S_\tau(u_n, h_n).$$

- $S_\tau$ is discontinuous.
- (A. Stuart & H. Lamba) There exists $\gamma \in (0, 1)$ and $C > 0$ such that
$$h \leqslant C \left( \frac{\tau}{\|f(u)\|} \right)^\gamma .$$
- (G. Hall & D.J. Higham) $\|f(u)\| \approx 0 \Rightarrow$ stability restricts the step-size.

# Stable Fixed Point Example

Consider the method RK1(2) applied to the linear system

$$\dot{u} = \begin{bmatrix} -5 & 0 \\ 0 & -1 \end{bmatrix}, \quad u = \begin{bmatrix} u_1 \\ u_2 \end{bmatrix}, \quad u(0) = [1, 10^{-4}]^T.$$




$(0, 0)$ – stable fixed point.

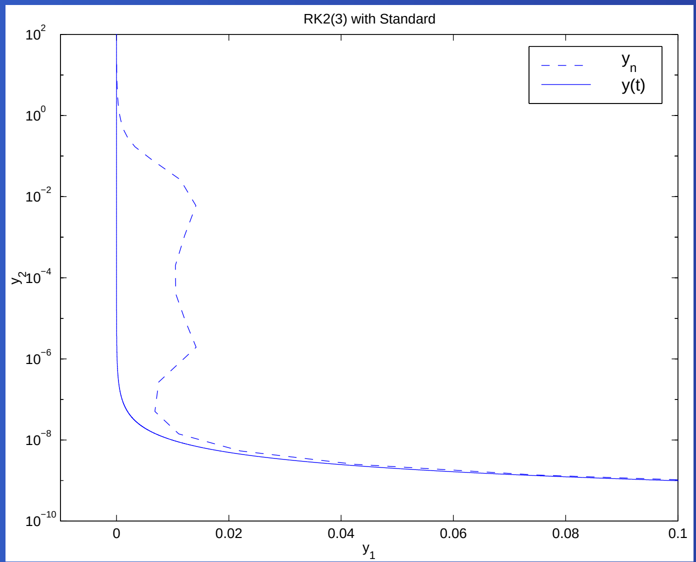
For this method, the numerical solution gives persistent spurious oscillations and the $y_1$ component has $\mathcal{O}(\tau)$ oscillation about the fixed point.

# Methods RK2(3) and RK4(5) applied to Saddle Point Example

$$\dot{u} = \begin{pmatrix} -1 & 0 \\ 0 & 1 \end{pmatrix} u, \quad u = \begin{pmatrix} u_1 \\ u_2 \end{pmatrix}, \quad u(0) = \left(0.99, 10^{-10}\right)^T.$$

# Methods RK2(3) and RK4(5) applied to Saddle Point Example

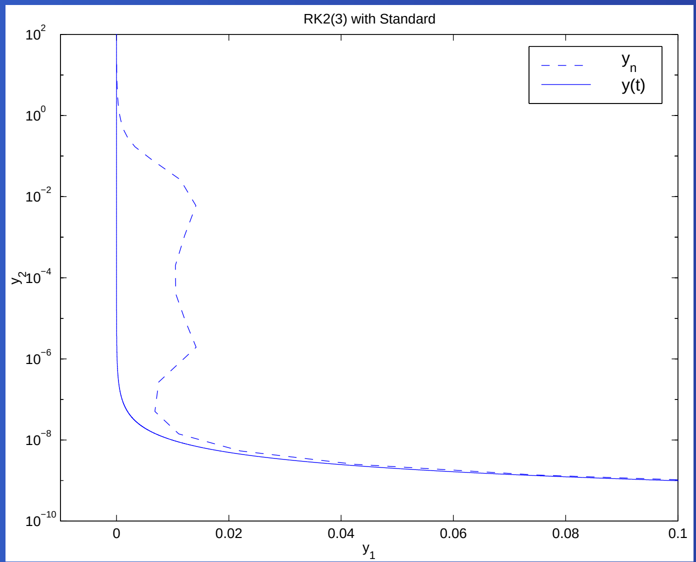
$$\dot{u} = \begin{pmatrix} -1 & 0 \\ 0 & 1 \end{pmatrix} u, \quad u = \begin{pmatrix} u_1 \\ u_2 \end{pmatrix}, \quad u(0) = (0.99, 10^{-10})^T.$$



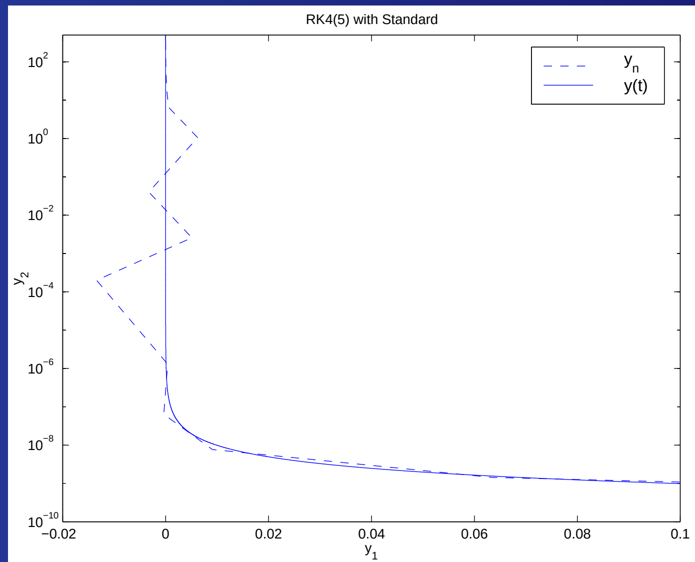

RK2(3) numerical solution does not pass close to fixed point or the local unstable manifold.

# Methods RK2(3) and RK4(5) applied to Saddle Point Example

$$\dot{u} = \begin{pmatrix} -1 & 0 \\ 0 & 1 \end{pmatrix} u, \quad u = \begin{pmatrix} u_1 \\ u_2 \end{pmatrix}, \quad u(0) = (0.99, 10^{-10})^T.$$




RK2(3) numerical solution does not pass close to fixed point or the local unstable manifold.

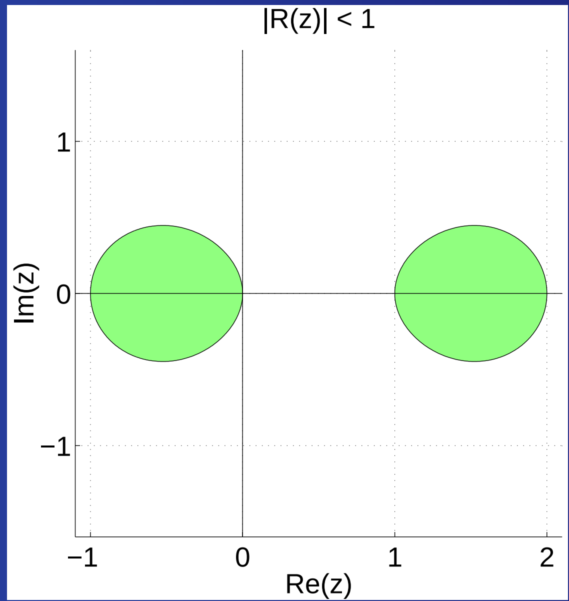



RK4(5) has spurious oscillations about the unstable manifold. Numerical solution can ultimately end up either side of the unstable manifold.

# Stable Saddle Point Example !

Consider 2-stage RK1(2) method with stability domain

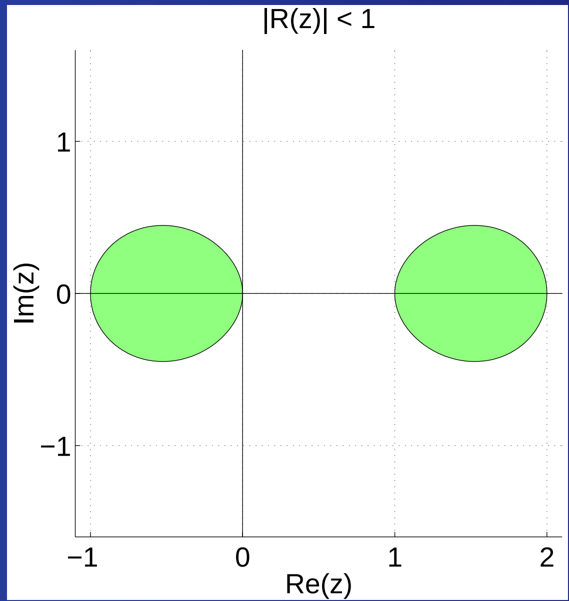
$$\begin{array}{c|cc} 0 & & \\ 1/2 & 1/2 & \\ \hline & 3 & 2 \\ & 0 & 1 \end{array}$$



# Stable Saddle Point Example !

Consider 2-stage RK1(2) method with stability domain

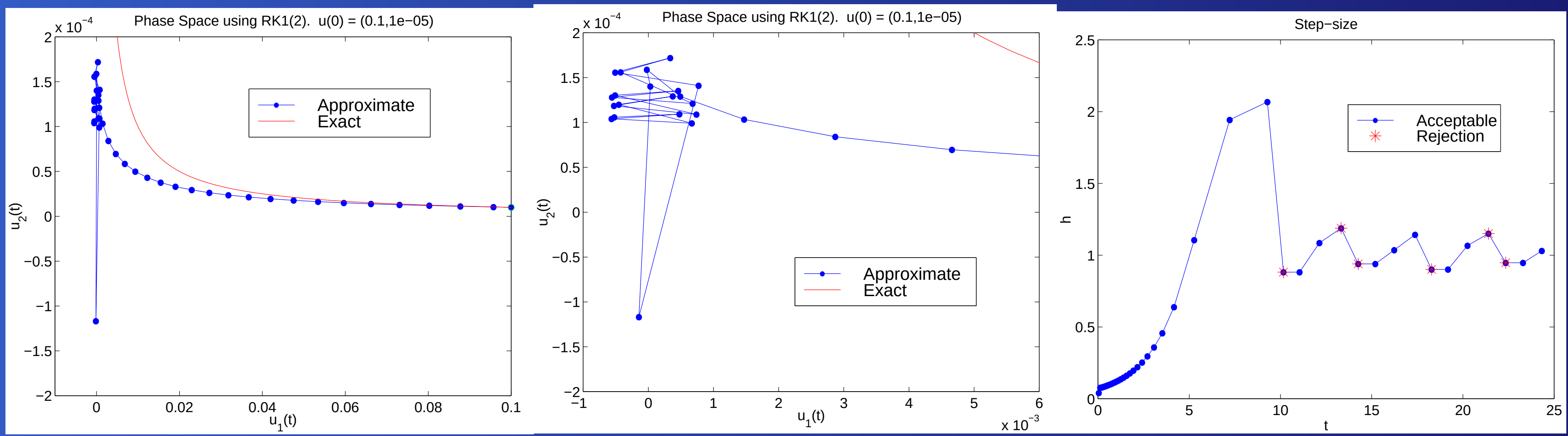
$$\begin{array}{c|cc} 0 & & \\ 1/2 & 1/2 & \\ \hline & 3 & 2 \\ & 0 & 1 \end{array}$$




Apply this method to

$$\dot{u} = \begin{pmatrix} -1 & 0 \\ 0 & 1 \end{pmatrix} u, \qquad u(0) = U \in \mathbb{R}^2.$$

# Stable Saddle Point Example !




In an $\mathcal{O}(\tau)$-neighbourhood of the origin, the step-size oscillates about 1.

Numerical solution near stable manifold becomes trapped near fixed point.

Spurious stable invariant object in numerical flow.

# Stable Saddle Point Example !

Stability function is $R(z) = 1 + z - z^2$. With $z = h\lambda$ and here $\lambda = \pm 1$. Consider fixed step-size.
For $h \in (0,1)$, the numerical manifolds are

$$W_h^s(0) = \{(x,y) \in \mathbb{R}^2 \mid y = 0\} \quad \text{and} \quad W_h^u(0) = \{(x,y) \in \mathbb{R}^2 \mid x = 0\}.$$

For $h \in (1,2)$, the numerical manifolds are

$$W_h^s(0) = \{(x,y) \in \mathbb{R}^2 \mid x = 0\} \quad \text{and} \quad W_h^u(0) = \{(x,y) \in \mathbb{R}^2 \mid y = 0\}.$$

## Stable Saddle Point Example !

Stability function is $R(z) = 1 + z - z^2$. With $z = h\lambda$ and here $\lambda = \pm 1$. Consider fixed step-size.

For $h \in (0, 1)$, the numerical manifolds are

$$W_h^s(0) = \{(x, y) \in \mathbb{R}^2 \mid y = 0\} \quad \text{and} \quad W_h^u(0) = \{(x, y) \in \mathbb{R}^2 \mid x = 0\}.$$

For $h \in (1, 2)$, the numerical manifolds are

$$W_h^s(0) = \{(x, y) \in \mathbb{R}^2 \mid x = 0\} \quad \text{and} \quad W_h^u(0) = \{(x, y) \in \mathbb{R}^2 \mid y = 0\}.$$

When $h$ crosses 1, the manifolds are reversed.
In adaptive algorithm this creates a chaotic attractor which persists for all $\tau > 0$.
Important to keep the step-size below linear (un)stability limits.

McGill

LaTeX

# Approximation of Local Unstable Manifolds with a Variable Step-Size Runge-Kutta Pair

Let $\sigma > 0$. The local unstable set of $\widehat{u}$, $W_\tau^{u,\sigma}(\widehat{u})$, is the set of $(u,h)$ such that there exists a backward orbit under $S_\tau$

$$\{(u_{-n}, h_{-n})\}_{n=0}^{\infty} \subset \mathbb{R}^d \times (0, h_{\max}],$$

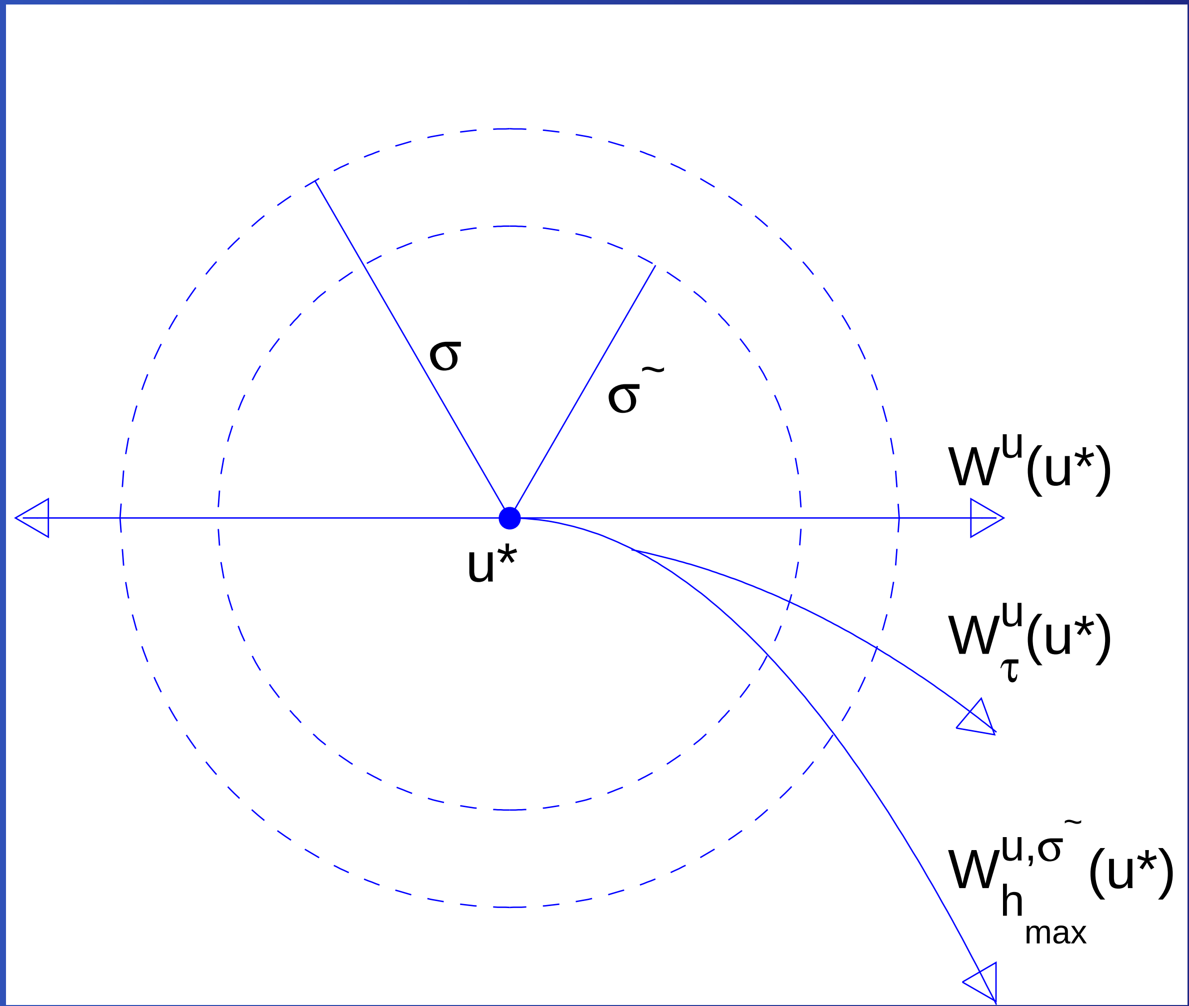
such that $(u,h) = (u_0, h_0)$, $\|u_{-n} - \widehat{u}\| \leqslant \sigma \ \forall n \geqslant 0$, and $u_{-n} \to \widehat{u}$ as $n \to \infty$.

Let $\sigma \geqslant \delta > 0$. $W_{\tau,\delta}^{u,\sigma}(\widehat{u})$ is the set of $(u,h)$ such that there exists a finite backward orbit under $S_\tau$

$$\{(u_{-n}, h_{-n})\}_{n=0}^{N} \subset \mathbb{R}^d \times (0, h_{\max}],$$

such that $(u,h) = (u_0, h_0)$, $\|u_{-n} - \widehat{u}\| \leqslant \sigma \quad \forall n = 0, \dots, N$, and $u_{-N} \in W_{h_{\max}}^{u,\delta}(\widehat{u})$.

# Approximation of Local Unstable Manifolds with a Variable Step-Size Runge-Kutta Pair



# An adaptive "Stable Manifold" Theorem

Lemma There exists $\delta = \mathcal{O}(\tau)$ and $h_{\max}$ sufficiently small and independent of $\tau$ such that

$$W^{u,\sigma}_{\tau,\delta}(\widehat{u}) = W^{u,\sigma}_{\tau}(\widehat{u}).$$

Theorem Apply an RK$p(\widetilde{p})$ method with $|p-\widetilde{p}| = 1$ to $\dot{u} = f(u)$. Let $\widehat{u}$ be a hyperbolic equilibrium and $f \in C^{\max\{p,\widetilde{p}\}}(\mathbb{R}^d)$. Then $\exists \sigma^+, H^+ > 0$ such that for $h_{\max} \in (0, H^+)$ & $\sigma \in (0, \sigma^+)$

$$d_H\left(W^{u,\sigma}(\widehat{u}),\ \mathcal{P}_u W^{u,\sigma}_{\tau}(\widehat{u})\right) \to 0 \qquad \text{as } \tau \to 0$$

where $\mathcal{P}_u : (u,h) \in \mathbb{R}^{d+1} \to u \in \mathbb{R}^d$ is the projection operator.

That is, the local unstable manifolds of the dynamical system and the unstable set of the RK$p(\widetilde{p})$ are close for small $\tau$.

# Phase Space Stability Error Control

Standard algorithm performs well during finite-time integration with fixed initial condition.

# Phase Space Stability Error Control

Standard algorithm performs well during finite-time integration with fixed initial condition.
However unless $h_{\max}$ is less than the linear stability limit the algorithm

- admits spurious fixed points;

# Phase Space Stability Error Control

Standard algorithm performs well during finite-time integration with fixed initial condition.
However unless $h_{\max}$ is less than the linear stability limit the algorithm

- admits spurious fixed points;
- performs badly around a stable fixed point;

# Phase Space Stability Error Control

Standard algorithm performs well during finite-time integration with fixed initial condition.
However unless $h_{\max}$ is less than the linear stability limit the algorithm

- admits spurious fixed points;
- performs badly around a stable fixed point;
- performs badly near saddle points.

McGill

LaTeX

# Phase Space Stability Error Control

Standard algorithm performs well during finite-time integration with fixed initial condition.
However unless $h_{\max}$ is less than the linear stability limit the algorithm

- admits spurious fixed points;
- performs badly around a stable fixed point;
- performs badly near saddle points.

Don't want to restrict $h_{\max}$ on whole phase space because of poor behaviour near fixed points

# Phase Space Stability Error Control

Standard algorithm performs well during finite-time integration with fixed initial condition.
However unless $h_{\max}$ is less than the linear stability limit the algorithm

- admits spurious fixed points;
- performs badly around a stable fixed point;
- performs badly near saddle points.

Don't want to restrict $h_{\max}$ on whole phase space because of poor behaviour near fixed points
Don't want to introduce expensive algorithm to compute linear stability limit near fixed points.

# Phase Space Stability Error Control

Standard algorithm performs well during finite-time integration with fixed initial condition.
However unless $h_{\max}$ is less than the linear stability limit the algorithm

- admits spurious fixed points;
- performs badly around a stable fixed point;
- performs badly near saddle points.

Introduce new phase space based error control to automatically control the step-size relative to the stability limit.

# Phase Space (PS$_\theta$) Error Control

We demand at each step the phase space (PS$_\theta$) error control

$$\begin{aligned}\|u_{n+1} - u_n - h_n[(1-\theta)f(u_n) + \theta f(u_{n+1})]\| \\ \leq \quad \varphi h_n \|(1-\theta)f(u_n) + \theta f(u_{n+1})\|, \quad \varphi \in (0,1).\end{aligned}$$

# Phase Space (PS$_\theta$) Error Control

We demand at each step the phase space (PS$_\theta$) error control

$$\begin{aligned}&\|u_{n+1} - u_n - h_n[(1-\theta)f(u_n) + \theta f(u_{n+1})]\| \\ &\qquad \leq \quad \varphi h_n \|(1-\theta)f(u_n) + \theta f(u_{n+1})\|, \quad \varphi \in (0,1).\end{aligned}$$

$h_n\|(1-\theta)f(u_n) + \theta f(u_{n+1})\|$ is approximation to arc length evolved over step.

McGill

LaTeX

# Phase Space ($PS_\theta$) Error Control

We demand at each step the phase space ($PS_\theta$) error control

$$\begin{aligned}&\|u_{n+1} - u_n - h_n[(1-\theta)f(u_n) + \theta f(u_{n+1})]\| \\ &\qquad \le \quad \varphi h_n\|(1-\theta)f(u_n) + \theta f(u_{n+1})\|, \quad \varphi \in (0,1).\end{aligned}$$

$h_n\|(1-\theta)f(u_n) + \theta f(u_{n+1})\|$ is approximation to arc length evolved over step.
So $PS_\theta$ error control bounds an approximation to local error by a fraction $\varphi$ of an approximation to solution arc length in phase space. So is a phase space error control.

# Phase Space ($PS_\theta$) Error Control

We demand at each step the phase space ($PS_\theta$) error control

$$\begin{aligned}\|u_{n+1} - u_n - h_n[(1-\theta)f(u_n) + \theta f(u_{n+1})]\| \\ \leq \quad \varphi h_n\|(1-\theta)f(u_n) + \theta f(u_{n+1})\|, \quad \varphi \in (0,1).\end{aligned}$$

$h_n\|(1-\theta)f(u_n) + \theta f(u_{n+1})\|$ is approximation to arc length evolved over step.
So $PS_\theta$ error control bounds an approximation to local error by a fraction $\varphi$ of an approximation to solution arc length in phase space. So is a phase space error control.

Will show it also acts as a stability control.

# Phase Space ($PS_\theta$) Error Control

We demand at each step the phase space ($PS_\theta$) error control

$$\begin{aligned} \|u_{n+1} - u_n - h_n[(1-\theta)f(u_n) + \theta f(u_{n+1})]\| \\ \leq \quad \varphi h_n \|(1-\theta)f(u_n) + \theta f(u_{n+1})\|, \quad \varphi \in (0,1). \end{aligned}$$

$h_n\|(1-\theta)f(u_n)+\theta f(u_{n+1})\|$ is approximation to arc length evolved over step.
So $PS_\theta$ error control bounds an approximation to local error by a fraction $\varphi$ of an approximation to solution arc length in phase space. So is a phase space error control.

Will show it also acts as a stability control.

Will combine this error control with standard error control; and demand both are satisfied at every step.

# Key Features of this Error Control

- Negligible additional computation is needed;

## Key Features of this Error Control

- Negligible additional computation is needed;
- Away from fixed points the standard error control is sufficient to ensure that the $PS_\theta$ condition is satisfied.

# Key Features of this Error Control

- Negligible additional computation is needed;
- Away from fixed points the standard error control is sufficient to ensure that the $PS_\theta$ condition is satisfied.
- Prevents spurious fixed points;

# Key Features of this Error Control

- Negligible additional computation is needed;
- Away from fixed points the standard error control is sufficient to ensure that the $\mathrm{PS}_\theta$ condition is satisfied.
- Prevents spurious fixed points;
- Forces convergence to stable fixed points;

# Key Features of this Error Control

- Negligible additional computation is needed;
- Away from fixed points the standard error control is sufficient to ensure that the PS$_\theta$ condition is satisfied.
- Prevents spurious fixed points;
- Forces convergence to stable fixed points;
- Gives stable step-size sequence with suitable step-size selection mechanism

## Key Features of this Error Control

- Negligible additional computation is needed;
- Away from fixed points the standard error control is sufficient to ensure that the $PS_\theta$ condition is satisfied.
- Prevents spurious fixed points;
- Forces convergence to stable fixed points;
- Gives stable step-size sequence with suitable step-size selection mechanism
- Good behaviour near saddle points

# Step-size selection

$$R(u_n, h_n) = \frac{\|u_{n+1} - u_n - h_n[(1-\theta)f(u_n) + \theta f(u_{n+1})]\|}{h_n\|(1-\theta)f(u_n) + \theta f(u_{n+1})\|} \leqslant \varphi.$$

Next step: $h^{\theta}_{n+1} = \left(\frac{\chi\varphi}{R(u_n,h_n)}\right)^{1/\tilde{q}} h_n,$ where by Taylor series

- Order $p \geq 2$ and $\theta \neq 1/2 \Rightarrow \tilde{q} = 1$;
- Order $p \geq 3$ and $\theta = 1/2 \Rightarrow \tilde{q} = 2$;

$\chi \in (0,1)$ is safety factor.

## Step-size selection

$$R(u_n, h_n) = \frac{\|u_{n+1} - u_n - h_n[(1-\theta)f(u_n) + \theta f(u_{n+1})]\|}{h_n\|(1-\theta)f(u_n) + \theta f(u_{n+1})\|} \leqslant \varphi.$$

Next step: $h^\theta_{n+1} = \left(\frac{\chi\varphi}{R(u_n,h_n)}\right)^{1/\tilde{q}} h_n,$ where by Taylor series

- Order $p \geq 2$ and $\theta \neq 1/2 \Rightarrow \tilde{q} = 1$;
- Order $p \geq 3$ and $\theta = 1/2 \Rightarrow \tilde{q} = 2$;

$\chi \in (0,1)$ is safety factor. New step-size selected as

$$h_{n+1} = \min\left[h^s_{n+1},\, h^\theta_{n+1},\, \alpha h_n\right],$$

where $h^s_{n+1}$ given by standard time-stepping strategy.
$\alpha > 1$ is a maximum step-size ratio , $\alpha = 5$.

## Linear system

Theorem Consider forward Euler method under $\mathrm{PS}_\theta$ error control in $\|\bullet\|_\infty$ with $\varphi \leqslant \theta/(1-\theta)$ applied to

$$u_t = \Lambda u, \quad \Lambda = \mathrm{Diag}[\lambda_1, \lambda_2, \cdots, \lambda_d], \quad u(0) = u_0 \in \mathbb{R}^d$$

where $\lambda_1 < \lambda_2 < \cdots < \lambda_d < 0$. Then $\|u^n\| \to 0$ as $n \to \infty$ with:

McGill

LaTeX

# Linear system

Theorem Consider forward Euler method under PS$_\theta$ error control in $\| \bullet \|_\infty$ with $\varphi \leqslant \theta/(1-\theta)$ applied to

$$u_t = \Lambda u, \quad \Lambda = \mathrm{Diag}[\lambda_1, \lambda_2, \cdots, \lambda_d], \quad u(0) = u_0 \in \mathbb{R}^d$$

where $\lambda_1 < \lambda_2 < \cdots < \lambda_d < 0$. Then $\|u^n\| \to 0$ as $n \to \infty$ with:

1. $u_d^n \to 0$ mono$^{ty}$ as $n \to \infty$;

McGill

LaTeX

## Linear system

Theorem Consider forward Euler method under PS$_\theta$ error control in $\| \bullet \|_\infty$ with $\varphi \leqslant \theta/(1-\theta)$ applied to

$$u_t = \Lambda u, \quad \Lambda = \mathrm{Diag}[\lambda_1, \lambda_2, \cdots, \lambda_d], \quad u(0) = u_0 \in \mathbb{R}^d$$

where $\lambda_1 < \lambda_2 < \cdots < \lambda_d < 0$. Then $\|u^n\| \to 0$ as $n \to \infty$ with:

1. $u_d^n \to 0$ mono$^{ty}$ as $n \to \infty$;
2. If $\lambda_i \geq \frac{\theta(1+\varphi)}{\varphi}\lambda_d$, then $u_i^n \to 0$ & $\frac{u_i^n}{u_d^n} \to 0$ mono$^{ty}$ as $n \to \infty$;

## Linear system

Theorem Consider forward Euler method under $\mathrm{PS}_\theta$ error control in $\|\bullet\|_\infty$ with $\varphi \leqslant \theta/(1-\theta)$ applied to

$$u_t = \Lambda u, \quad \Lambda = \mathrm{Diag}[\lambda_1, \lambda_2, \cdots, \lambda_d], \quad u(0) = u_0 \in \mathbb{R}^d$$

where $\lambda_1 < \lambda_2 < \cdots < \lambda_d < 0$. Then $\|u^n\| \to 0$ as $n \to \infty$ with:

1. $u_d^n \to 0$ mono$^{ty}$ as $n \to \infty$;
2. If $\lambda_i \geq \frac{\theta(1+\varphi)}{\varphi}\lambda_d$, then $u_i^n \to 0$ & $\frac{u_i^n}{u_d^n} \to 0$ mono$^{ty}$ as $n \to \infty$;
3. If $\frac{\theta(1+\varphi)}{\varphi}\lambda_d > \lambda_i \geq \left[\frac{2\theta(1+\varphi)}{\varphi} - 1\right]\lambda_d$, then $u_i^n \to 0$ & $\frac{|u_i^n|}{|u_d^n|} \to 0$;

# Linear system

Theorem Consider forward Euler method under $\mathrm{PS}_\theta$ error control in $\|\bullet\|_\infty$ with $\varphi \leqslant \theta/(1-\theta)$ applied to

$$u_t = \Lambda u, \quad \Lambda = \mathrm{Diag}[\lambda_1, \lambda_2, \cdots, \lambda_d], \quad u(0) = u_0 \in \mathbb{R}^d$$

where $\lambda_1 < \lambda_2 < \cdots < \lambda_d < 0$. Then $\|u^n\| \to 0$ as $n \to \infty$ with:

1. $u_d^n \to 0$ mono$^{ty}$ as $n \to \infty$;
2. If $\lambda_i \geq \frac{\theta(1+\varphi)}{\varphi}\lambda_d$, then $u_i^n \to 0$ & $\frac{u_i^n}{u_d^n} \to 0$ mono$^{ty}$ as $n \to \infty$;
3. If $\frac{\theta(1+\varphi)}{\varphi}\lambda_d > \lambda_i \geq \left[\frac{2\theta(1+\varphi)}{\varphi} - 1\right]\lambda_d$, then $u_i^n \to 0$ & $\frac{|u_i^n|}{|u_d^n|} \to 0$;
4. Otherwise $u_i^n \to 0$ as $n \to \infty$ with $\limsup_{n\to\infty} \frac{|u_i^n|}{|u_d^n|} < \frac{\varphi}{\theta - \frac{\varphi}{1+\varphi}}$;

## Linear system

Theorem Consider forward Euler method under PS$_\theta$ error control in $\|\bullet\|_\infty$ with $\varphi \leqslant \theta/(1-\theta)$ applied to

$$u_t = \Lambda u, \quad \Lambda = \mathrm{Diag}[\lambda_1, \lambda_2, \cdots, \lambda_d], \quad u(0) = u_0 \in \mathbb{R}^d$$

where $\lambda_1 < \lambda_2 < \cdots < \lambda_d < 0$. Then $\|u^n\| \to 0$ as $n \to \infty$ with:

1. $u_d^n \to 0$ mono$^{ty}$ as $n \to \infty$;
2. If $\lambda_i \geq \frac{\theta(1+\varphi)}{\varphi}\lambda_d$, then $u_i^n \to 0$ & $\frac{u_i^n}{u_d^n} \to 0$ mono$^{ty}$ as $n \to \infty$;
3. If $\frac{\theta(1+\varphi)}{\varphi}\lambda_d > \lambda_i \geq \left[\frac{2\theta(1+\varphi)}{\varphi} - 1\right]\lambda_d$, then $u_i^n \to 0$ & $\frac{|u_i^n|}{|u_d^n|} \to 0$;
4. Otherwise $u_i^n \to 0$ as $n \to \infty$ with $\limsup_{n\to\infty} \frac{|u_i^n|}{|u_d^n|} < \frac{\varphi}{\theta - \frac{\varphi}{1+\varphi}}$;
5. Let $\theta_n$ be angle between $u^n$ and $[0, \ldots, 0, 1] \in \mathbb{R}^d$. If $u_d^0 \neq 0$

$$\liminf_{n\to\infty} \cos\theta_n \geqslant 1 - \frac{1}{2}(d-1)\frac{\varphi^2}{\theta^2} + \mathcal{O}(\varphi^3).$$

# Remarks

- Can extend to arbitrary norms.

## Remarks

- Can extend to arbitrary norms.
- Can extend to $\dot{u} = Au$ (i.e. non-diagonal). and nonlinear hyperbolic equilibria.

McGill

LaTeX

## Remarks

- Can extend to arbitrary norms.
- Can extend to $\dot{u} = Au$ (i.e. non-diagonal). and nonlinear hyperbolic equilibria.
- Extends to non-stiff saddle points.

# Remarks

- Can extend to arbitrary norms.
- Can extend to $\dot{u} = Au$ (i.e. non-diagonal). and nonlinear hyperbolic equilibria.
- Extends to non-stiff saddle points.
- Can extend to arbitrary methods.

# Remarks

- Can extend to arbitrary norms.
- Can extend to $\dot{u} = Au$ (i.e. non-diagonal). and nonlinear hyperbolic equilibria.
- Extends to non-stiff saddle points.
- Can extend to arbitrary methods.
- Bound in 4 independent of stiffness/eigenvalues and can be made arbitray small by reducing $\varphi$.

# Remarks

- Can extend to arbitrary norms.
- Can extend to $\dot{u} = Au$ (i.e. non-diagonal). and nonlinear hyperbolic equilibria.
- Extends to non-stiff saddle points.
- Can extend to arbitrary methods.
- Bound in 4 independent of stiffness/eigenvalues and can be made arbitray small by reducing $\varphi$.
- The exact solution is tangent to $[0, 0, \ldots, 0, 1]$ at fixed point, so 5 gives bound on angle between exact and numerical solutions at fixed point. Reducing $\varphi$ makes angle arbitrarily small (independent of the stiffness/eigenvalues).

McGill

LaTeX

## Proof

Forward Euler method gives

$$u^{n+1} = R(h_n A)u^n = \mathsf{Diag}[1 + h_n\lambda_1, \cdots, 1 + h_n\lambda_d]u^n.$$

# Proof

Forward Euler method gives

$$u^{n+1} = R(h_n A)u^n = \text{Diag}[1 + h_n\lambda_1, \cdots, 1 + h_n\lambda_d]u^n.$$

With $\infty$-norm, PS$_\theta$ error control becomes

$$\left\| \begin{bmatrix} \theta h_n^2 \lambda_1^2 u_1^n \\ \theta h_n^2 \lambda_2^2 u_2^n \\ \vdots \\ \theta h_n^2 \lambda_d^2 u_d^n \end{bmatrix} \right\|_\infty \leqslant \varphi h_n \left\| \begin{bmatrix} \lambda_1(1 + \theta\lambda_1 h_n)u_1^n \\ \lambda_2(1 + \theta\lambda_2 h_n)u_2^n \\ \vdots \\ \lambda_d(1 + \theta\lambda_d h_n)u_d^n \end{bmatrix} \right\|_\infty .$$

Note Unlike standard control not trivially true at near point.

# Proof

Forward Euler method gives

$$u^{n+1} = R(h_n A)u^n = \text{Diag}[1 + h_n\lambda_1, \cdots, 1 + h_n\lambda_d]u^n.$$

With $\infty$-norm, PS$_\theta$ error control becomes

$$\left\| \begin{bmatrix} \theta h_n^2 \lambda_1^2 u_1^n \\ \theta h_n^2 \lambda_2^2 u_2^n \\ \vdots \\ \theta h_n^2 \lambda_d^2 u_d^n \end{bmatrix} \right\|_\infty \leqslant \varphi h_n \left\| \begin{bmatrix} \lambda_1(1 + \theta\lambda_1 h_n)u_1^n \\ \lambda_2(1 + \theta\lambda_2 h_n)u_2^n \\ \vdots \\ \lambda_d(1 + \theta\lambda_d h_n)u_d^n \end{bmatrix} \right\|_\infty .$$

hence for some $i \in \{1, 2, \ldots, d\}$

$$h_n \theta \lambda_i^2 |u_i^n| \leqslant -\varphi\lambda_i |1 + \theta\lambda_i h_n||u_i^n|.$$

# Proof

$$h_n \theta \lambda_i^2 |u_i^n| \leqslant -\varphi \lambda_i |1 + \theta \lambda_i h_n| |u_i^n|.$$

True if and only if

$$h_n \leqslant -\frac{\varphi}{\lambda_i \theta (1+\varphi)} \leqslant -\frac{\varphi}{\lambda_d \theta (1+\varphi)}.$$

# Proof

$$h_n \theta \lambda_i^2 |u_i^n| \leqslant -\varphi \lambda_i |1 + \theta \lambda_i h_n| |u_i^n|.$$

True if and only if

$$h_n \leqslant -\frac{\varphi}{\lambda_i \theta (1+\varphi)} \leqslant -\frac{\varphi}{\lambda_d \theta (1+\varphi)}.$$

So the $d^{th}$ condition always holds and monotonic convergence in this component follows.

## Proof

$$h_n \theta \lambda_i^2 |u_i^n| \leqslant -\varphi \lambda_i |1 + \theta \lambda_i h_n| |u_i^n|.$$

True if and only if

$$h_n \leqslant -\frac{\varphi}{\lambda_i \theta (1+\varphi)} \leqslant -\frac{\varphi}{\lambda_d \theta (1+\varphi)}.$$

So the $d^{th}$ condition always holds and monotonic convergence in this component follows.

Prove 2 and 3 by showing that $d^{th}$ condition implies (monotonic) convergence for these components.

# Proof

$$h_n\theta\lambda_i^2|u_i^n| \leqslant -\varphi\lambda_i|1+\theta\lambda_i h_n||u_i^n|.$$

True if and only if

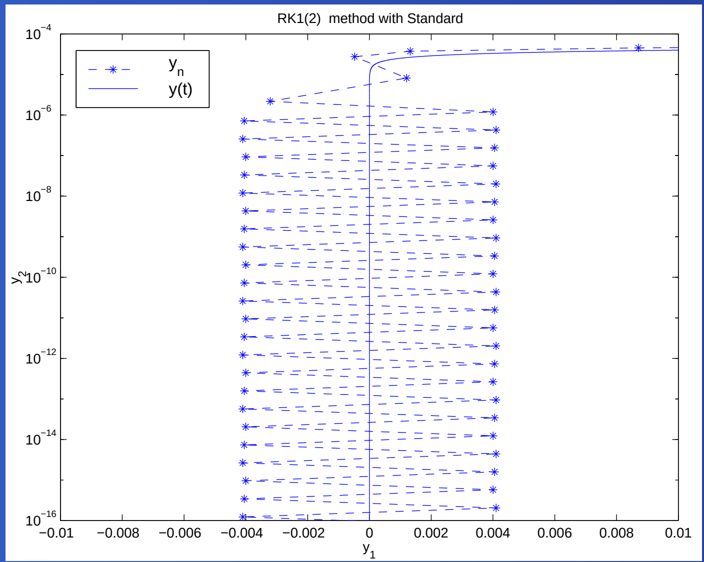
$$h_n \leqslant -\frac{\varphi}{\lambda_i\theta(1+\varphi)} \leqslant -\frac{\varphi}{\lambda_d\theta(1+\varphi)}.$$

So the $d^{th}$ condition always holds and monotonic convergence in this component follows.

Prove 2 and 3 by showing that $d^{th}$ condition implies (monotonic) convergence for these components.

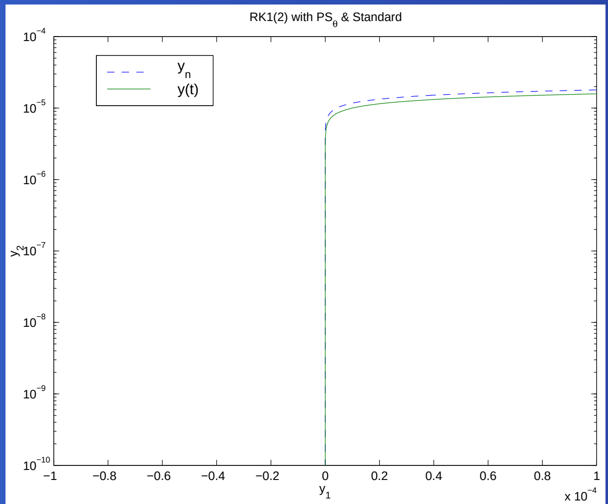
Prove 4 by showing that failure of the $i$-th condition bounds $|y_i^n/y_d^n|$ and hard work. 5 follows on noting that in limit $n \to \infty$ all components bounded in terms of $d^{th}$ component. $\square$

# Stable Fixed Point Example (Revisited)



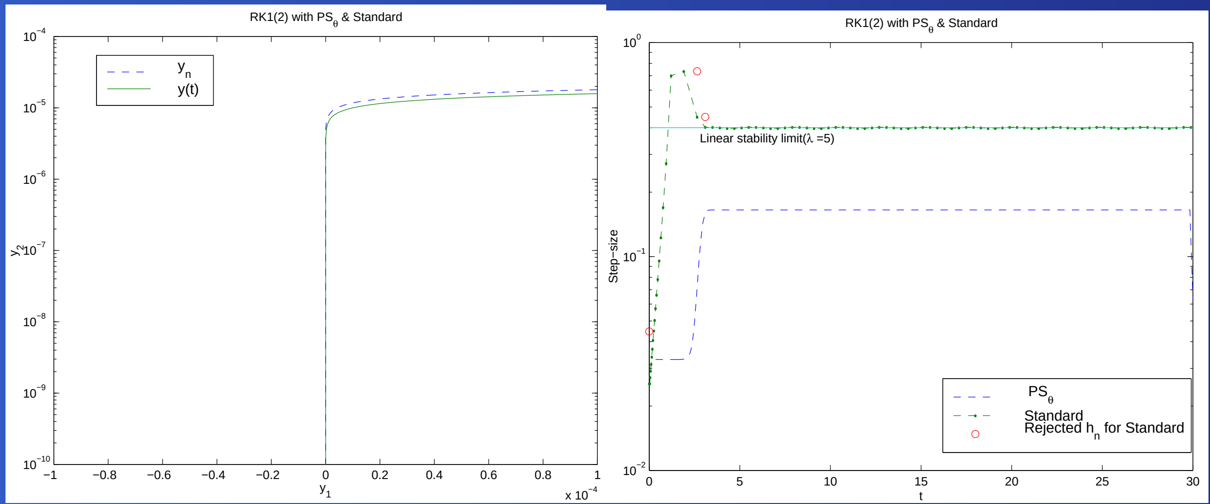

Recall solution with standard algorithm

# Stable Fixed Point Example (Revisited)




With PS$_\theta$ spurious oscillation is removed

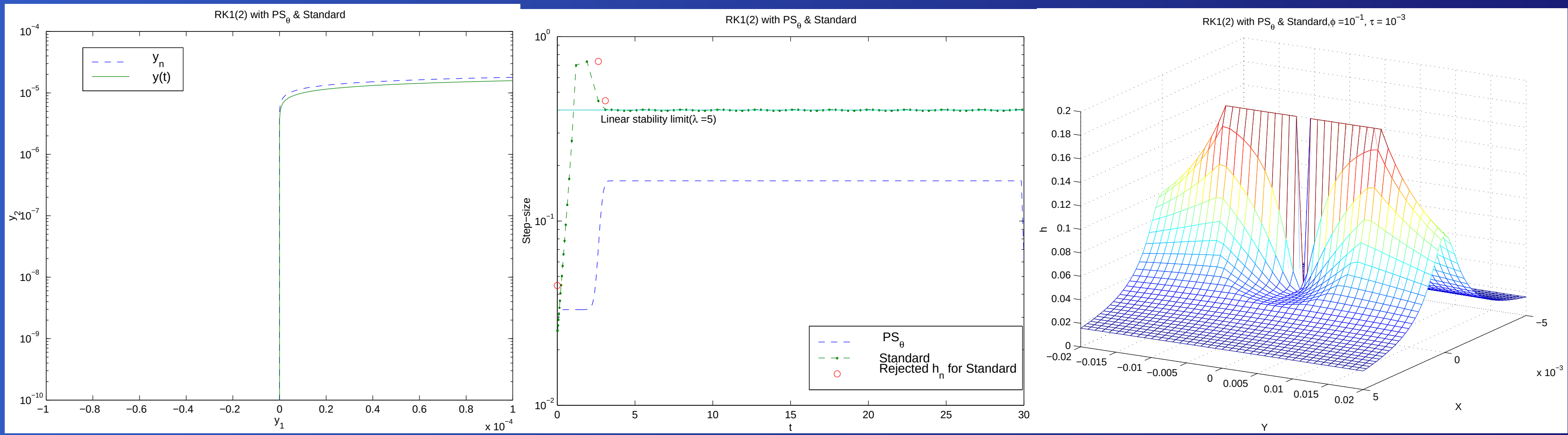
# Stable Fixed Point Example (Revisited)




With PS$_\theta$ spurious oscillation is removed
Step-size is kept below stability limit.

McGill

# Stable Fixed Point Example (Revisited)



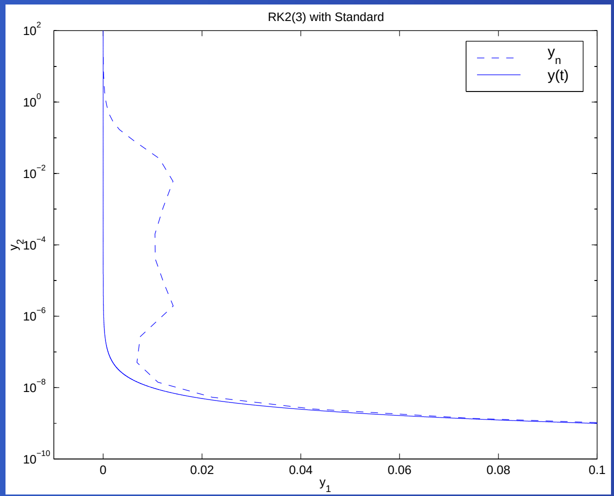

With $PS_\theta$ spurious oscillation is removed

Step-size is kept below stability limit.

Step-sizes bounded near fixed point.

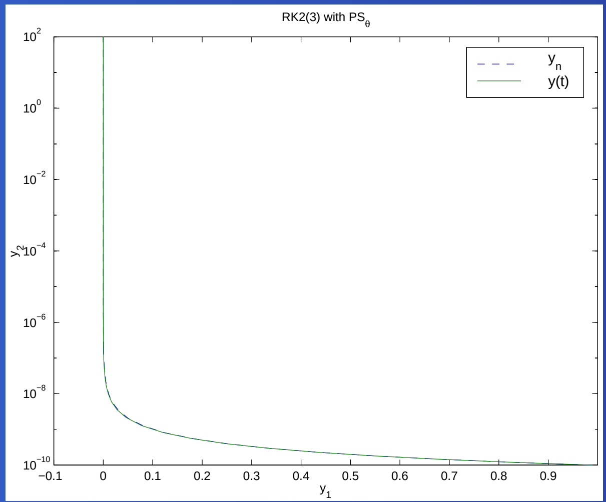
$PS_\theta$ only determines step-size near fixed point.

# Saddle Point Example (Revisited)



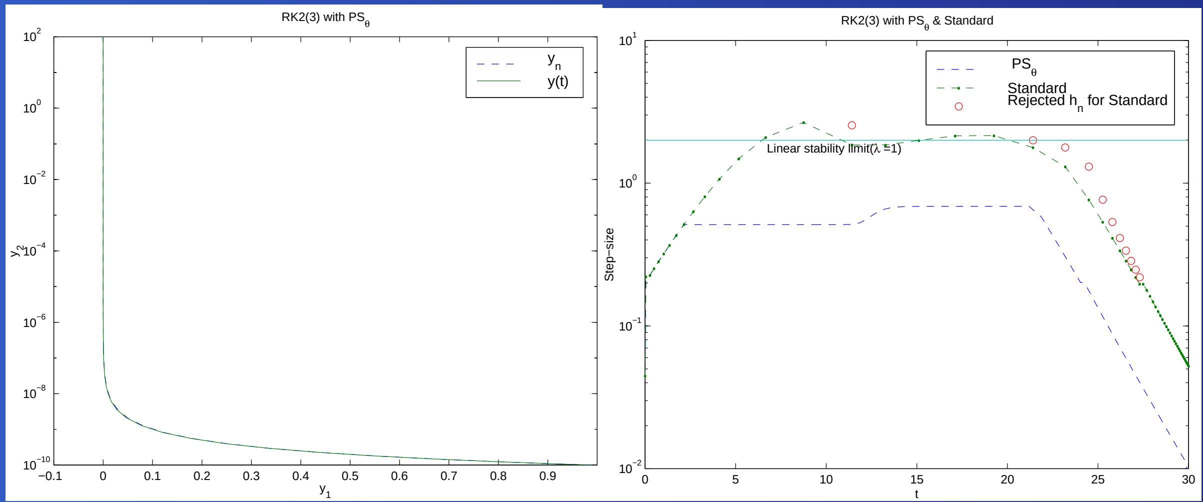

Recall solution with RK2(3) standard algorithm

# Saddle Point Example (Revisited)



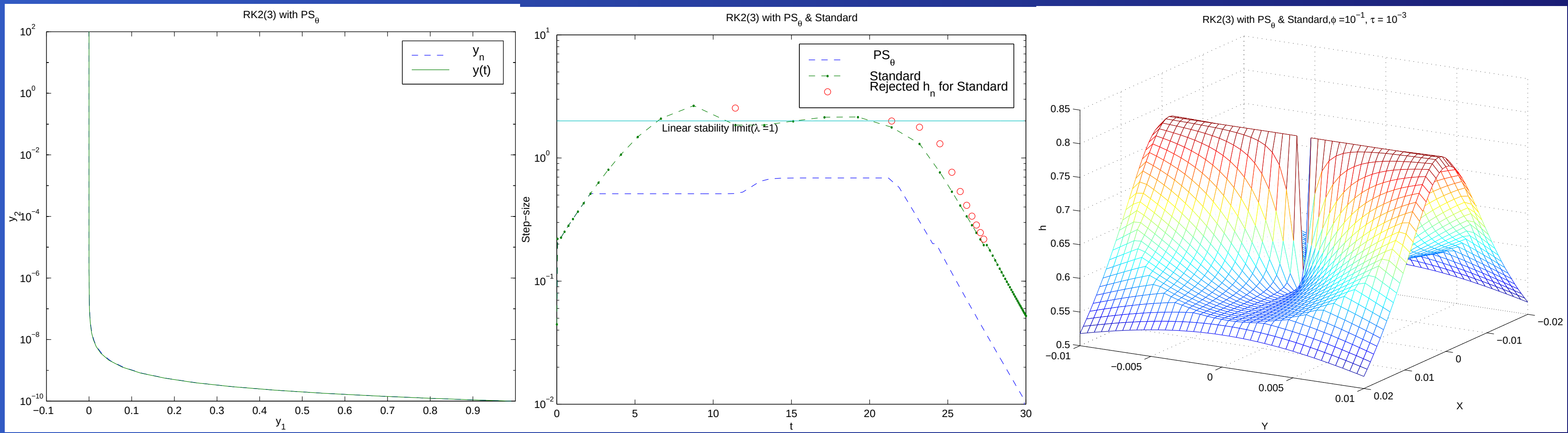

With $PS_\theta$ spurious oscillation is removed

# Saddle Point Example (Revisited)




With PS$_\theta$ spurious oscillation is removed
Step-size is kept below stability limit.

# Saddle Point Example (Revisited)



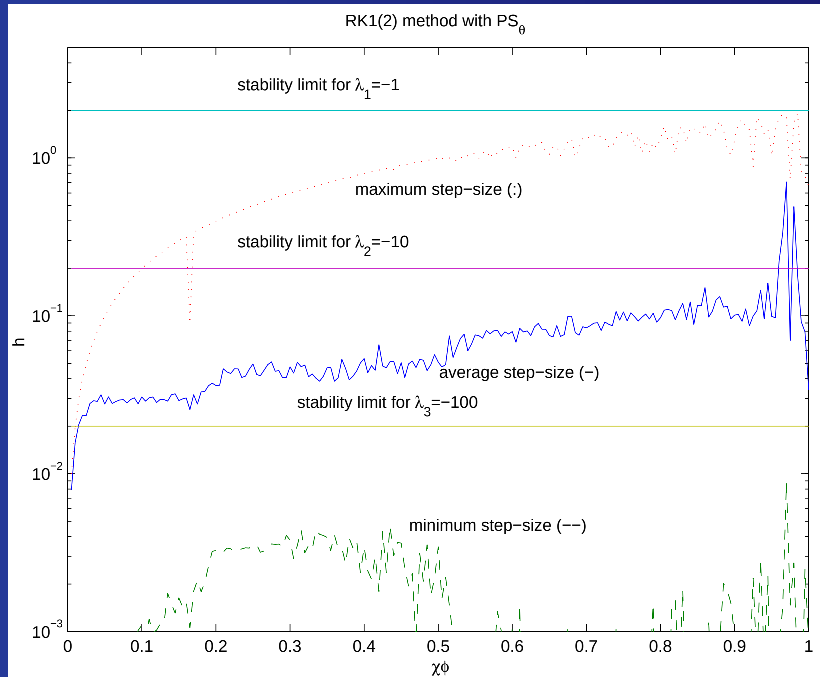

With PS$_\theta$ spurious oscillation is removed
Step-size is kept below stability limit.
Step-sizes bounded near fixed point.
PS$_\theta$ only determines step-size near fixed point.

# Large Average Step-Sizes

RK1(2) Method applied to

$$\dot{u} = \begin{pmatrix} -1 & 0 & 0 \\ 0 & -10 & 0 \\ 0 & 0 & -100 \end{pmatrix} u.$$

Step-size oscillates.



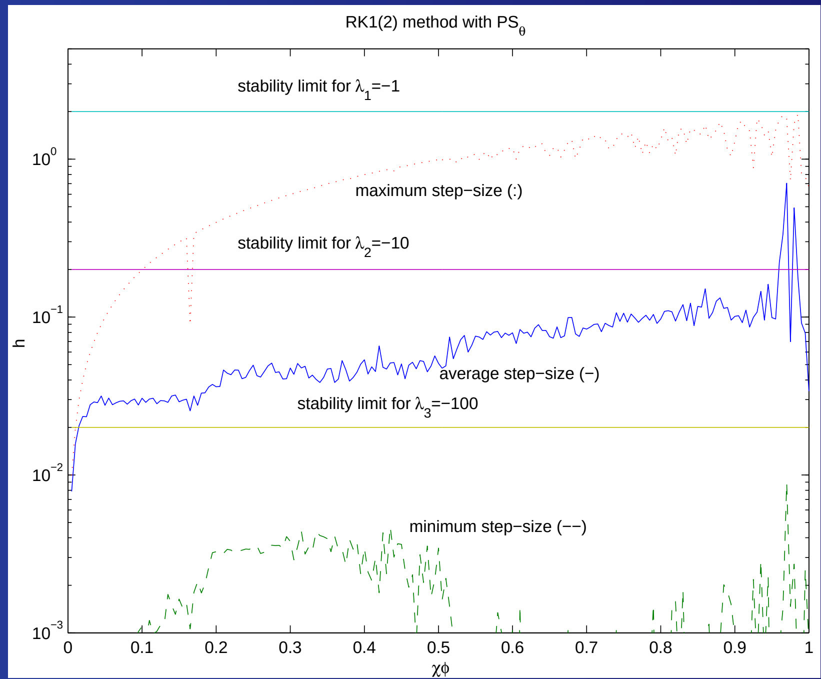
# Large Average Step-Sizes

RK1(2) Method applied to

$$\dot{u} = \begin{pmatrix} -1 & 0 & 0 \\ 0 & -10 & 0 \\ 0 & 0 & -100 \end{pmatrix} u.$$

Step-size oscillates.

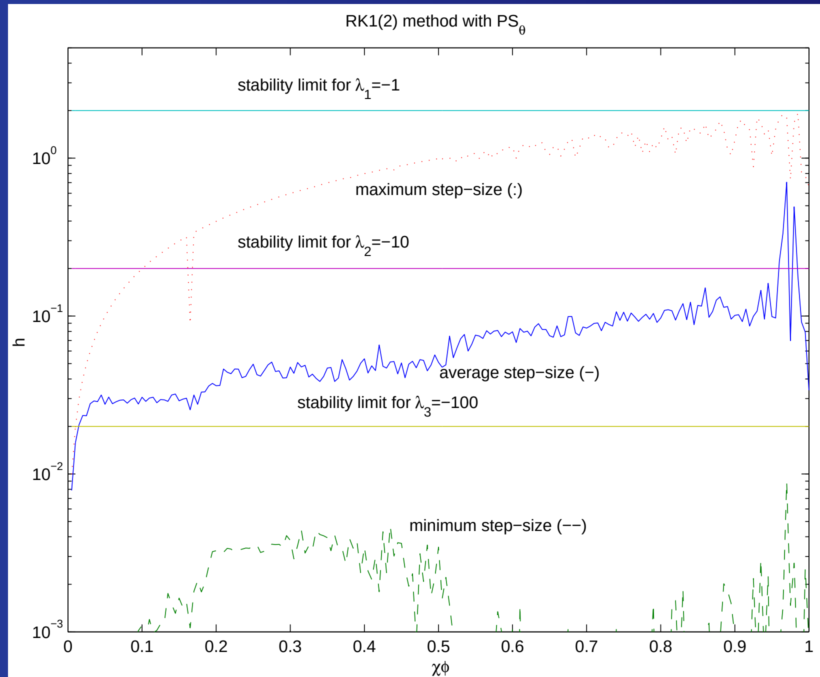
All steps below $\lambda = -1$ stability limit; monotonic convergence of this component.



# Large Average Step-Sizes

RK1(2) Method applied to

$$\dot{u} = \begin{pmatrix} -1 & 0 & 0 \\ 0 & -10 & 0 \\ 0 & 0 & -100 \end{pmatrix} u.$$




Step-size oscillates.

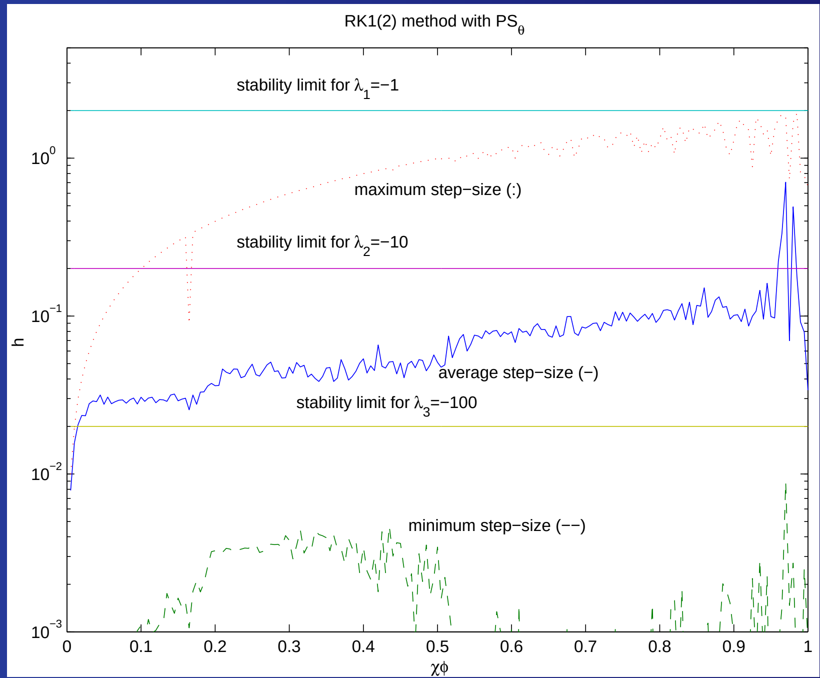
All steps below $\lambda = -1$ stability limit; monotonic convergence of this component.

Average step-size below $\lambda = -10$ stability limit, but for $\varphi > 0.1$ some step-sizes above this limit.

# Large Average Step-Sizes

RK1(2) Method applied to

$$\dot{u} = \begin{pmatrix} -1 & 0 & 0 \\ 0 & -10 & 0 \\ 0 & 0 & -100 \end{pmatrix} u.$$




Step-size oscillates.

All steps below $\lambda = -1$ stability limit; monotonic convergence of this component.

Average step-size below $\lambda = -10$ stability limit, but for $\varphi > 0.1$ some step-sizes above this limit.

Except for $\varphi$ tiny, average and maximum step-sizes above $\lambda = -100$ stability limit. But convergence to fixed point enforced.

# Conclusions

Standard algorithm behaves poorly near saddle points.
Stiff methods do not resolve problem for saddles.

# Conclusions

Standard algorithm behaves poorly near saddle points.
Stiff methods do not resolve problem for saddles.

$\mathrm{PS}_\theta$

- proved to give correct behaviour near stable fixed points
- gives correct behaviour near non-stiff saddles

## Conclusions

Standard algorithm behaves poorly near saddle points.
Stiff methods do not resolve problem for saddles.

$\mathrm{PS}_\theta$

- proved to give correct behaviour near stable fixed points
- gives correct behaviour near non-stiff saddles

Ongoing

- $\mathrm{PS}_\theta$ currently being implemented in standard ODE solver