Multirate explicit stabilized method in mixed precision arithmetic

Giacomo Rosilho de Souza
Euler Institute, Lugano, Switzerland

Joint work with
Matteo Croci
Oden Institute, Austin, USA

CAIMS 2023 - Fredericton

Table of Contents

Motivating Mixed-Precision Arithmetic

Mixed-precision algorithms combine low- and high-precision computations in order to benefit from:

- Performance, energy, and memory gains of low-precision,
- Accuracy of high-precision.

Format	unit roundoff u
bfloat16 (half)	$2^{-8} \approx 3.91 \times 10^{-3}$
fp16 (half)	$2^{-11} \approx 4.88 \times 10^{-4}$
fp32 (single)	$2^{-24} \approx 5.96 \times 10^{-8}$
fp64 (double)	$2^{-53} \approx 1.11 \times 10^{-16}$

Trend: roundoff unit u is getting larger!!!

Mixed precision integration

Assumptions, computational setup, and notation:

- Computations in high-precision arithmetic are assumed to be exact.
- u is the roundoff unit of the low-precision format: $u \approx 10^{-3}$.
- Round to nearest.
- Computations performed in low precision are denoted by a hat $\widehat{\ }$, the error model is:

$$\widehat{a \operatorname{op} b} = (1+\delta)(a \operatorname{op} b), \qquad |\delta| < u, \qquad \operatorname{op} \in \{+, -, *, /\},$$

Remember: $\widehat{\ }$ produces a relative error $\approx u \approx 10^{-3}$.

Mixed precision integration

Motivating example on a linear problem

Consider

$$y' = Ay, \qquad y(0) = y_0$$

and the integrators

For accuracy

$$y_1 = y_0 + \Delta t A y_0,$$

$$y_1 = y_0 + \Delta t A y_0 + \frac{1}{2}\Delta t^2 A^2 y_0,$$

$$y_1 = y_0 + \Delta t A y_0 + c\Delta t^2 A^2 y_0 \,.$$

For stability

Goal: design mixed-precision versions of these integrators preserving the original accuracy.

For $y_1 = y_0 + \Delta t A y_0$:

Try 1:

$$\begin{aligned} \widehat{y}_1 &= \widehat{y_0 + \Delta t A y_0} = (1+\delta)(y_0 + \Delta t A y_0) \\ &= y_0 + \Delta t A y_0 + \mathcal{O}(u)\,. \end{aligned}$$

Local error: $\mathcal{O}(u)$.

Global error: $\mathcal{O}(u\Delta t^{-1})$.

Divergence

Mixed precision integration

Motivating example on a linear problem

Consider

$$y' = Ay, \qquad y(0) = y_0$$

and the integrators

$$y_1 = y_0 + \Delta t A y_0,$$
$$y_1 = y_0 + \Delta t A y_0 + \frac{1}{2}\Delta t^2 A^2 y_0,$$
$$y_1 = y_0 + \Delta t A y_0 + c\Delta t^2 A^2 y_0\,.$$

Goal: design mixed-precision versions of these integrators preserving the original accuracy.

For $y_1 = y_0 + \Delta t A y_0$:

Try 2:

$$\hat{y}_1 = y_0 + \Delta t \widehat{A y_0} = y_0 + \Delta t A y_0 + \mathcal{O}(\Delta t u \|A\| \|y_0\|)$$
$$= y_0 + \Delta t A y_0 + \mathcal{O}(\Delta t u)\,.$$

Local error: $\mathcal{O}(\Delta t u)$.

Global error: $\mathcal{O}(u)$.

Saturation

Mixed precision integration

Motivating example on a linear problem

Consider

$$y' = Ay, \qquad y(0) = y_0$$

and the integrators

$$y_1 = y_0 + \Delta t A y_0,$$

$$y_1 = y_0 + \Delta t A y_0 + \frac{1}{2}\Delta t^2 A^2 y_0,$$

$$y_1 = y_0 + \Delta t A y_0 + c\Delta t^2 A^2 y_0 \,.$$

Goal: design mixed-precision versions of these integrators preserving the original accuracy.

For $y_1 = y_0 + \Delta t A y_0 + \frac{1}{2}\Delta t^2 A^2 y_0$:

For accuracy

Try 3:

$$\widehat{y}_1 = y_0 + \Delta t A y_0 + \frac{1}{2}\Delta t^2 \widehat{A^2 y_0}$$

$$= y_0 + \Delta t A y_0 + \frac{1}{2}\Delta t^2 A^2 y_0 + \mathcal{O}(\Delta t^2 u) \,.$$

Local error: $\mathcal{O}(\Delta t^2 u)$.

Global error: $\mathcal{O}(\Delta t u)$.

Order reduction

Mixed-precision integration

Motivating example on a linear problem

Consider

$$y' = Ay, \qquad y(0) = y_0$$

and the integrators

$$y_1 = y_0 + \Delta t A y_0,$$
$$y_1 = y_0 + \Delta t A y_0 + \frac{1}{2}\Delta t^2 A^2 y_0,$$
$$y_1 = y_0 + \Delta t A y_0 + c\Delta t^2 A^2 y_0 \,.$$

Goal: design mixed-precision versions of these integrators preserving the original accuracy.

For $y_1 = y_0 + \Delta t A y_0 + c\Delta t^2 A^2 y_0$, $c \neq 1/2$:

Try 4:

For stability

$$\begin{aligned}\widehat{y}_1 &= y_0 + \Delta t A y_0 + c\Delta t^2 \widehat{A^2 y_0} \\ &= y_0 + \Delta t A y_0 + c\Delta t^2 A^2 y_0 + \mathcal{O}(\Delta t^2 u)\,.\end{aligned}$$

Local error: $\mathcal{O}(\Delta t^2 u)$.

Global error: $\mathcal{O}(\Delta t u)$.

Same order of convergence

Conclusion: harder to work with methods where coefficients are optimized for accuracy. But we can play with the stabilization terms.

Introduction to explicit stabilized methods

We want to solve, for instance,

$$y' = \nabla \cdot (A(y)\,\nabla y) + f(y).$$

We typically have:

Standard explicit solver: $\Delta t \leq Ch^2$,

Implicit solver: solves nonlinear problem.

With explicit stabilized methods:

- No step size Δt restrictions,
- No linear systems to solve.

Some differences with respect to standard explicit methods:

- Adaptive in the number of stages s,
- Given an order p, use an increased number of stages $s \geq p$,
- Gained freedom is used to optimise in the stability direction,
- Stability domain grows as $O(s^2)$,
- Work load scales as $O(\sqrt{\rho}) = O(h^{-1})$, not as $O(\rho) = O(h^{-2})$.

The Runge-Kutta-Chebyshev method

Consider

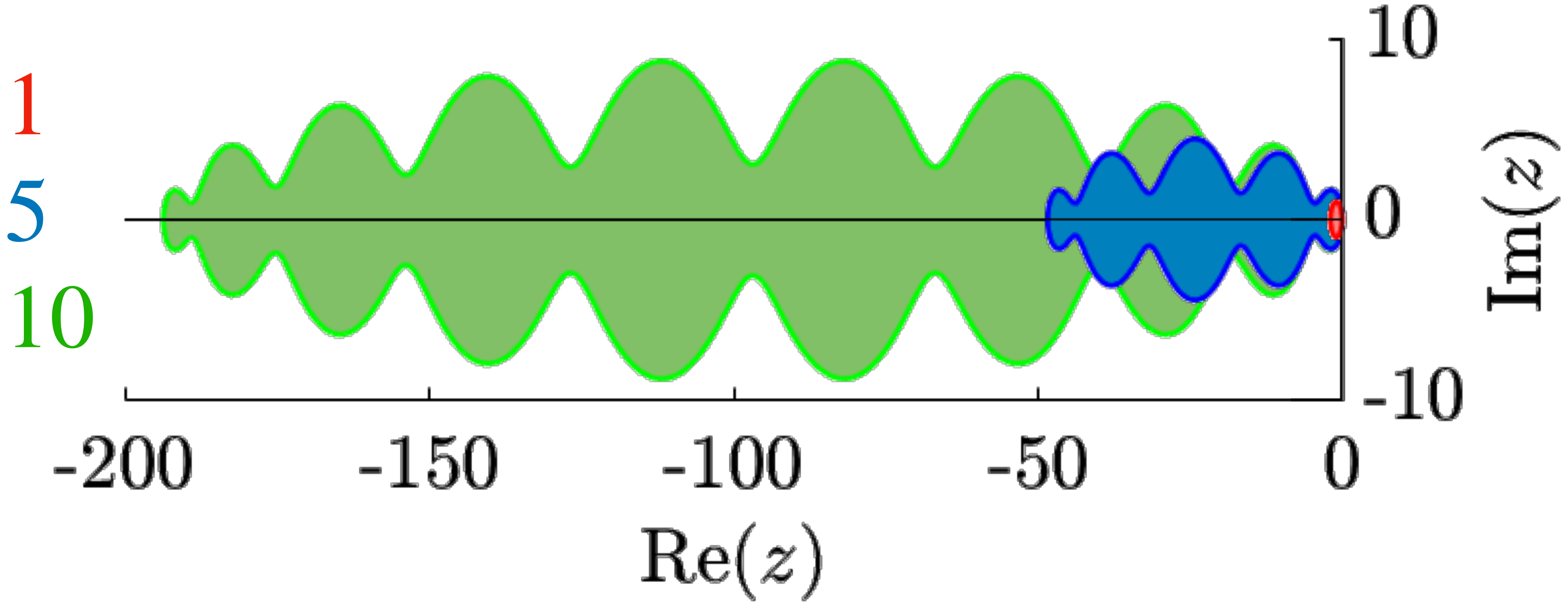
$$y' = f(y), \qquad y(0) = y_0 \,.$$

One step of RKC in δ-form is given by

$$d_0 = 0, \qquad d_1 = \mu_1 \Delta t\, f(y_0),$$
$$d_j = \nu_j\, d_{j-1} + \kappa_j\, d_{j-2} + \mu_j \Delta t\, f(y_0 + d_{j-1}), \quad j = 2, \ldots, s,$$
$$y_1 = y_0 + d_s,$$

with s satisfying $\Delta t \rho \le 2s^2$.

$s = 1$
$s = 5$
$s = 10$



- No step size restriction,
- Fully explicit,
- Straightforward to implement.

We note that the method needs:

- Only $p = 1{,}2$ function evaluations for accuracy,
- and $s - p$ for stability.

But every evaluation contributes to both, accuracy and stability! For a mixed-precision version, we need to refactor the method.

The Mixed-Precision Runge-Kutta-Chebyshev method

Original method:

$$d_0 = 0, \qquad d_1 = \mu_1 \Delta t\, f(y_0),$$
$$d_j = \nu_j\, d_{j-1} + \kappa_j\, d_{j-2} + \mu_j \Delta t\, f(y_0 + d_{j-1}), \quad j = 2, \ldots, s,$$
$$y_1 = y_0 + d_s,$$

Linearized method:

$$d_0 = 0, \qquad d_1 = \mu_1 \Delta t\, f(y_0),$$
$$d_j = \nu_j\, d_{j-1} + \kappa_j\, d_{j-2} + \mu_j \Delta t \left(f(y_0) + J(y_0) d_{j-1} \right)$$
$$y_1 = y_0 + d_s,$$

Mixed-precision method:

$$d_0 = 0, \qquad d_1 = \mu_1 \Delta t\, f(y_0),$$
$$d_j = \nu_j\, d_{j-1} + \kappa_j\, d_{j-2} + \mu_j \Delta t \left(f(y_0) + \widehat{J(y_0) d_{j-1}} \right)$$
$$y_1 = y_0 + d_s,$$

$\widehat{J(y_0) d_{j-1}}$ is computed with one low-precision evaluation of f.

Cost:

- 1 function evaluation in high-precision,
- $s - 1$ function evaluation in low-precision.

Low-precision Jacobian's computation

The mixed-precision RKC method is:

$$d_0 = 0, \qquad d_1 = \mu_1 \Delta t\, f(y_0), \qquad d_j = \nu_j\, d_{j-1} + \kappa_j\, d_{j-2} + \mu_j \Delta t \left(f(y_0) + \widehat{J(y_0) d_{j-1}} \right)$$

How do we approximate the Jacobian $\widehat{J(y_0) d_j}$ efficiently in low-precision?

Naive approach:

$$\begin{aligned} \widehat{J(y_0) d_j} &:= \hat{f}(y_0 + d_j) - f(y_0) = f(y_0 + d_j) - f(y_0) + \mathcal{O}(u) \\ &= J(y_0) d_j + \mathcal{O}(u + \|d_j\|^2) = J(y_0) d_j + \mathcal{O}(u + \Delta t^2) \end{aligned}$$

Local error: $\mathcal{O}(\Delta t u)$, Global error: $\mathcal{O}(u)$.

Low-precision Jacobian's computation

The mixed-precision RKC method is:

$$d_0 = 0, \qquad d_1 = \mu_1 \Delta t\, f(y_0), \qquad d_j = \nu_j\, d_{j-1} + \kappa_j\, d_{j-2} + \mu_j \Delta t \left(f(y_0) + \widehat{J(y_0) d_{j-1}} \right)$$

How do we approximate the Jacobian $\widehat{J(y_0)d_j}$ efficiently in low-precision?

Smarter approach:

$$\begin{aligned} \widehat{J(y_0)d_j} : &= \epsilon^{-1}\left(\hat{f}(y_0 + \epsilon d_j) - f(y_0)\right) = \epsilon^{-1}\left(f(y_0 + \epsilon d_j) - f(y_0) + \mathcal{O}(u)\right) \\ &= \epsilon^{-1}\left(J(y_0)\epsilon d_j + \mathcal{O}(u + \epsilon^2 \|d_j\|^2)\right) = J(y_0)d_j + \mathcal{O}(\epsilon^{-1}u + \epsilon \Delta t^2) \end{aligned}$$

Take $\epsilon = \sqrt{u}/\Delta t$, then $\mathcal{O}(\epsilon^{-1}u + \epsilon\Delta t^2) = \mathcal{O}(\Delta t\sqrt{u})$.

Local error: $\mathcal{O}(\Delta t^2 \sqrt{u})$, Global error: $\mathcal{O}(\Delta t\sqrt{u})$.

Convergence and stability

Convergence

The global error between the high-precision and the mixed-precision RKC method is[1]

$$\|y_n - \hat{y}_n\| = \mathcal{O}(\Delta t\sqrt{u})$$

A second-order scheme exists, with

$$\|y_n - \hat{y}_n\| = \mathcal{O}(\Delta t^2)$$

Stability

- Roundoff errors destroy any spectral relationship between the error term and the solution,
- A stability analysis in the classical sense is undoable,
- The best that we can do is a worst-case analysis that doesn't take into account roundoff errors' cancellation[1],
- Numerical experiments show that our mixed-precision schemes are stable[1].

[1] M. Croci, G. Rosilho de Souza, *Journal of Computational Physic, 464*, 2022.

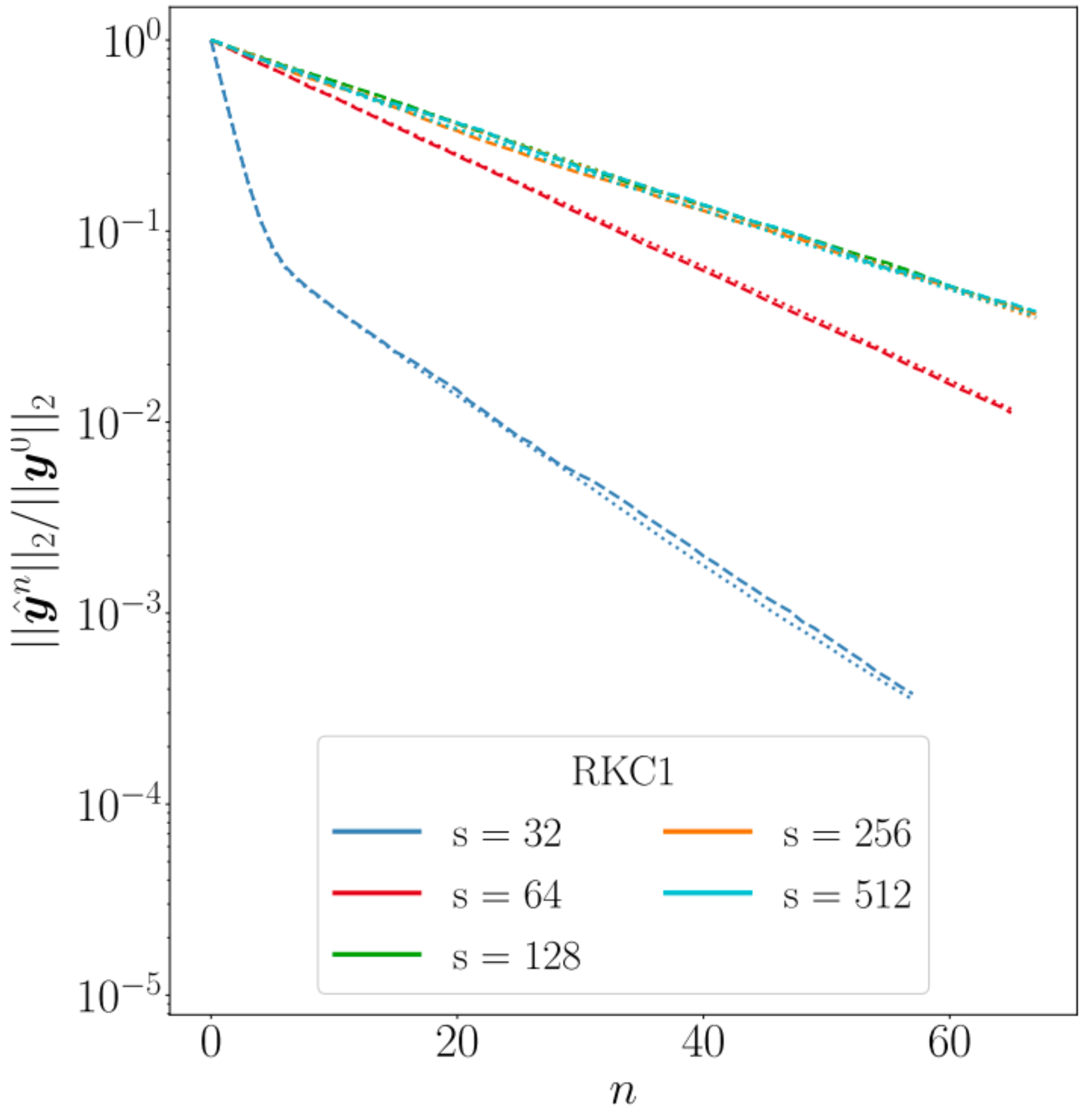
Stability Check

Solve

$$\begin{aligned} \frac{\partial u}{\partial t} &= 100\Delta u && \text{in } \Omega \times [0,T], \\ u(x,t) &= 0 && \text{on } \partial\Omega \times [0,T], \\ u(x,0) &= u_0(x) && \text{in } \Omega, \end{aligned}$$

with $\Omega = [0,1] \times [0,1]$, $T = 1$.

For different mesh sizes and fixed Δt, we check that the norm decreases in time.



Convergence experiment

Solve

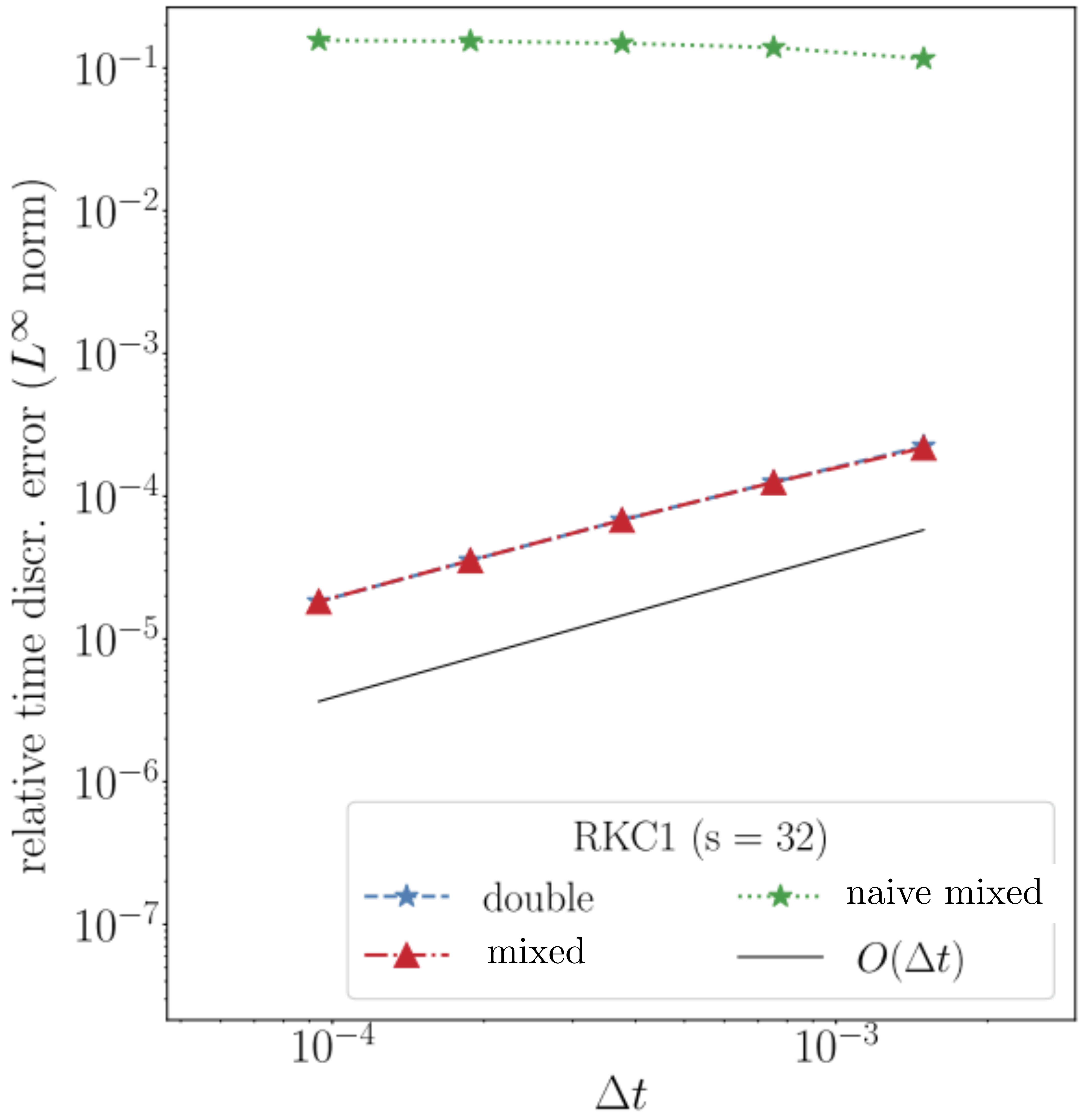
$$\begin{aligned} \frac{\partial u}{\partial t} &= \nabla \cdot (\|\nabla u\|_2^2 \nabla u) + f(x) && \text{in } \Omega \times [0,T], \\ u(x,t) &= 1 && \text{on } \partial\Omega \times [0,T], \\ u(x,0) &= 1 && \text{in } \Omega, \end{aligned}$$

with $\Omega = [0,1]$, $T = 1$.

For $h = 1/32 = 0.03125$ and fixed $s = 32$ we let $\Delta t \to 0$ and plot the errors

$$\frac{1}{u}\|\hat{u}_n - u(t_n)\|_{L^\infty((0,T),L^\infty(\Omega))}$$

For both RKC1 and RKC2.



Convergence experiment

Solve

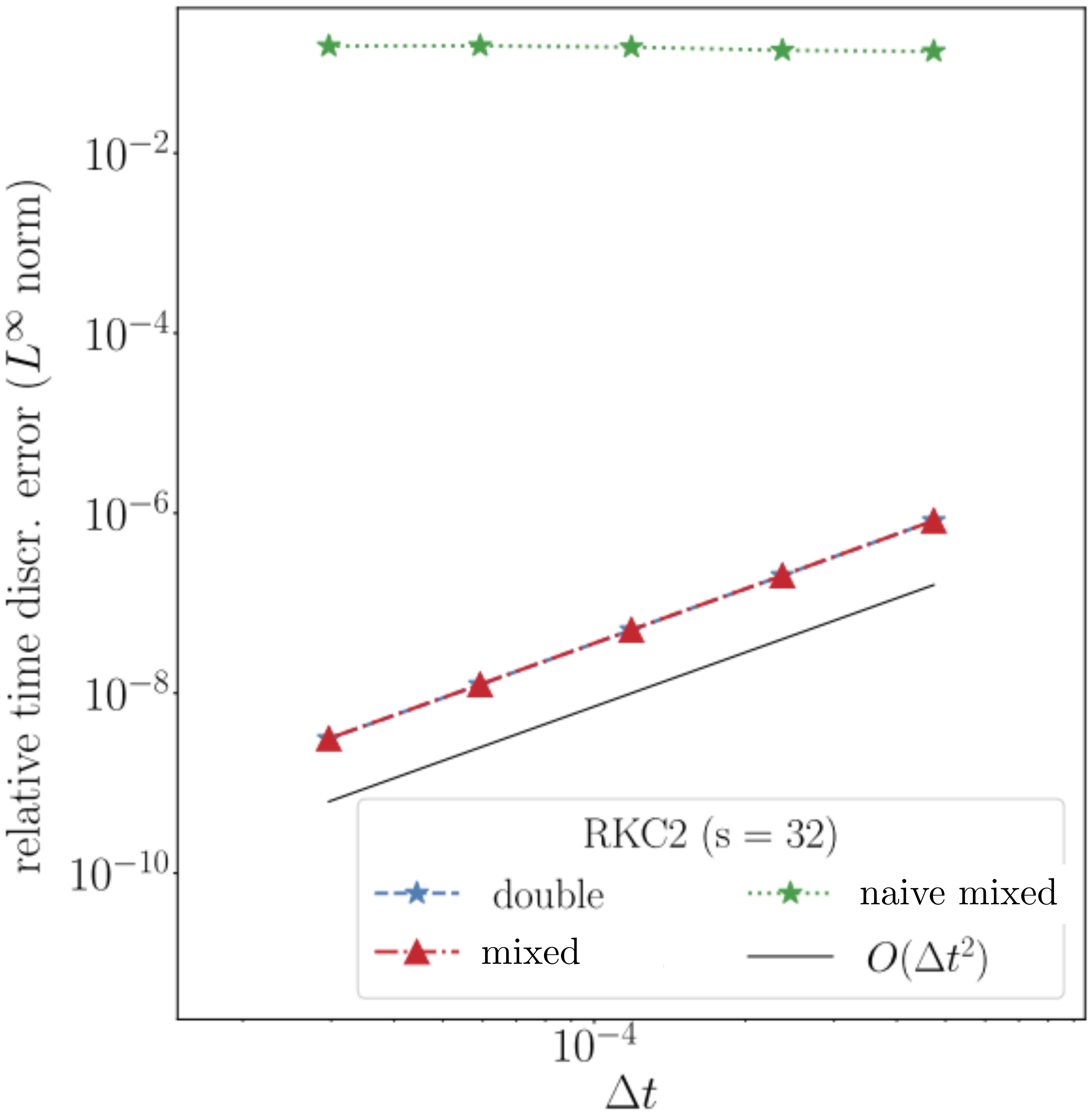
$$
\begin{aligned}
\frac{\partial u}{\partial t} &= \nabla \cdot (\|\nabla u\|_2^2 \nabla u) + f(x) && \text{in } \Omega \times [0,T], \\
u(x,t) &= 1 && \text{on } \partial\Omega \times [0,T], \\
u(x,0) &= 1 && \text{in } \Omega,
\end{aligned}
$$

with $\Omega = [0,1]$, $T = 1$.

For $h = 1/32 = 0.03125$ and fixed $s = 32$ we let $\Delta t \to 0$ and plot the errors

$$\frac{1}{u}\|\hat{u}_n - u(t_n)\|_{L^\infty((0,T),L^\infty(\Omega))}$$

For both RKC1 and RKC2.



Mixed precision multirate RKC method

Consider

$$y' = f_F(y) + f_S(y), \qquad y(0) = y_0,$$

with f_F stiff but cheap and f_S mildly stiff but expensive.

For RKC, number of expensive f_S evaluations is dictated by the few stiff terms in f_F and deteriorates efficiency.

We solve the *modified problem*

$$y_\eta' = f_\eta(y_\eta), \qquad y(0) = y_0,$$

With $\eta \geq 0$ a parameter used to tune the stiffness. For $\eta = 2/\rho_S$ the stiffness of f_η is same as f_S.

The *averaged force* is defined as

$$f_\eta(y) = \frac{1}{\eta}\left(u(\eta) - y\right)$$

With *auxiliary solution* u given by

$$u' = f_F(u) + f_S(y), \qquad u(0) = y\,.$$

The multirate RKC method is given by:

- Integrate $y_\eta' = f_\eta(y_\eta)$ with a RKC method.
- To evaluate f_η solve $u' = f_F(u) + f_S(y)$ with another RKC method.

Mixed precision multirate RKC method

The multirate RKC method:

$$d_0 = 0, \qquad d_1 = \mu_1 \Delta t\, \bar{f}_\eta(y_0),$$
$$d_j = \nu_j\, d_{j-1} + \kappa_j\, d_{j-2} + \mu_j \Delta t\, \bar{f}_\eta(y_0 + d_{j-1}),\ \ j = 2,\dots,s,$$
$$y_1 = y_0 + d_s,$$

With $\Delta t \rho_S \leq 2s^2$ and

$$h_0 = 0, \qquad h_1 = \alpha_1(f_F(y) + f_S(y)),$$
$$h_j = \beta_j h_{j-1} + \gamma_j h_{j-2} + \alpha_j(f_F(y + \eta h_{j-1}) + f_S(y)),$$
$$\bar{f}_\eta(y) = h_m,$$

Where $\eta \rho_F \leq 2m^2$, $\eta \approx \Delta t / s^2$. Cost is:

- s evaluations of f_S in high-precision,
- $s \cdot m$ evaluations of f_F in high-precision.

The mixed-precision multirate RKC method:

$$d_0 = 0, \qquad d_1 = \mu_1 \Delta t\, \hat{f}_\eta(y_0),$$
$$d_j = \nu_j\, d_{j-1} + \kappa_j\, d_{j-2} + \mu_j \Delta t \left(\hat{f}_\eta(y_0) + J_\eta\widehat{(y_0) d}_{j-1}\right)$$
$$y_1 = y_0 + d_s,$$

- $\hat{f}_\eta(y_0)$ computed applying a mixed-precision RKC method to $\bar{f}_\eta(y)$.
- $J_\eta\widehat{(y_0) d}_{j-1}$ computed applying a low-precision RKC method to $\bar{f}_\eta(y)$.
- 1 evaluation of f_F, f_S in high-precision,
- Remaining evaluations in low-precision.

Numerical experiments

Solve

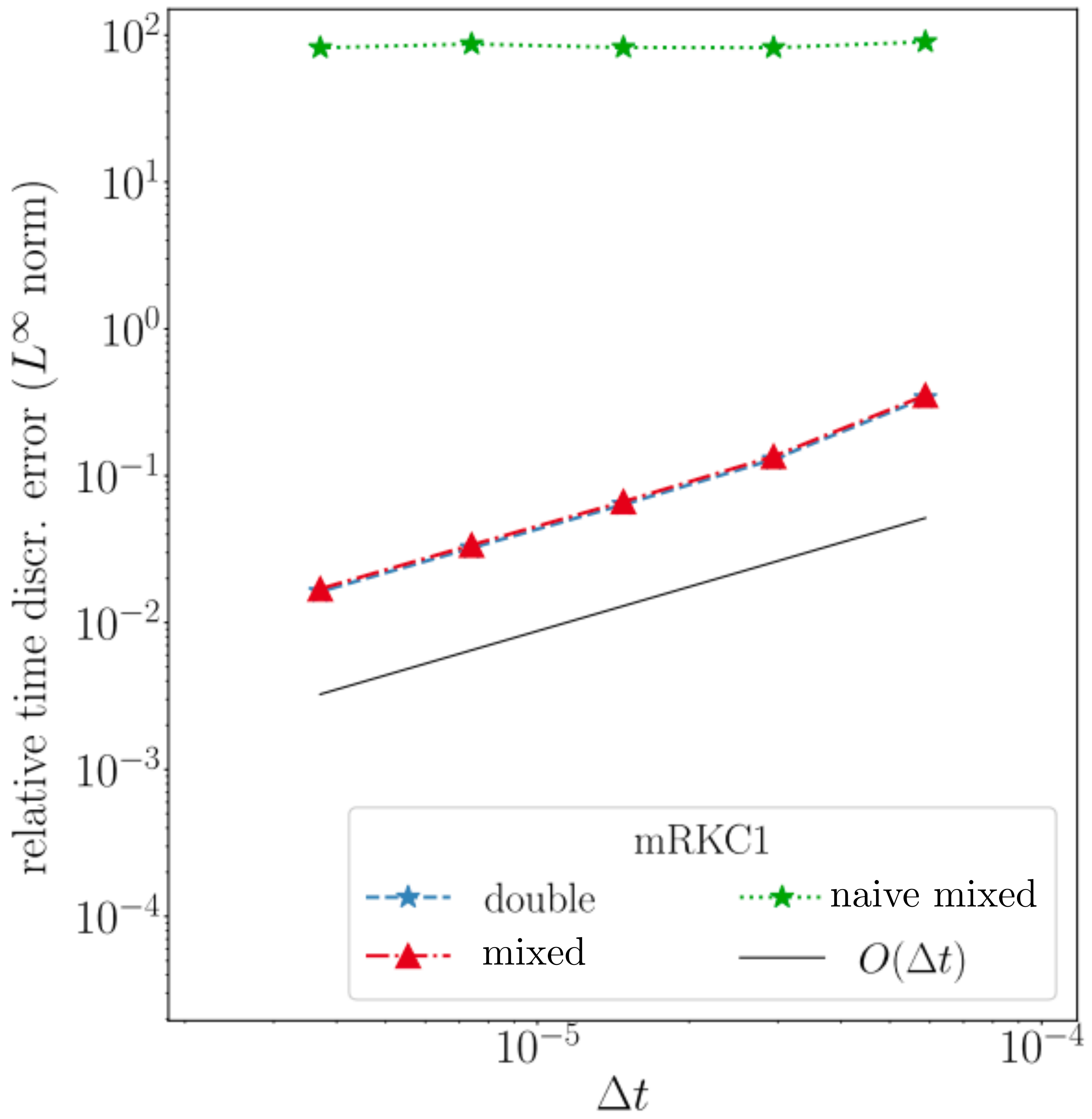
$$\begin{aligned} \frac{\partial u}{\partial t} &= 100\Delta u + f_S(u,x) && \text{in } \Omega \times [0,T], \\ u(x,t) &= 0 && \text{on } \partial\Omega \times [0,T], \\ u(x,0) &= u_0(x) && \text{in } \Omega, \end{aligned}$$

We fix mesh size $h = 0.0156$ and check convergence for $\Delta t \to 0$.

Plot errors

$$\frac{1}{u}\|\hat{u}_n - u(t_n)\| \text{ VS } \Delta t.$$

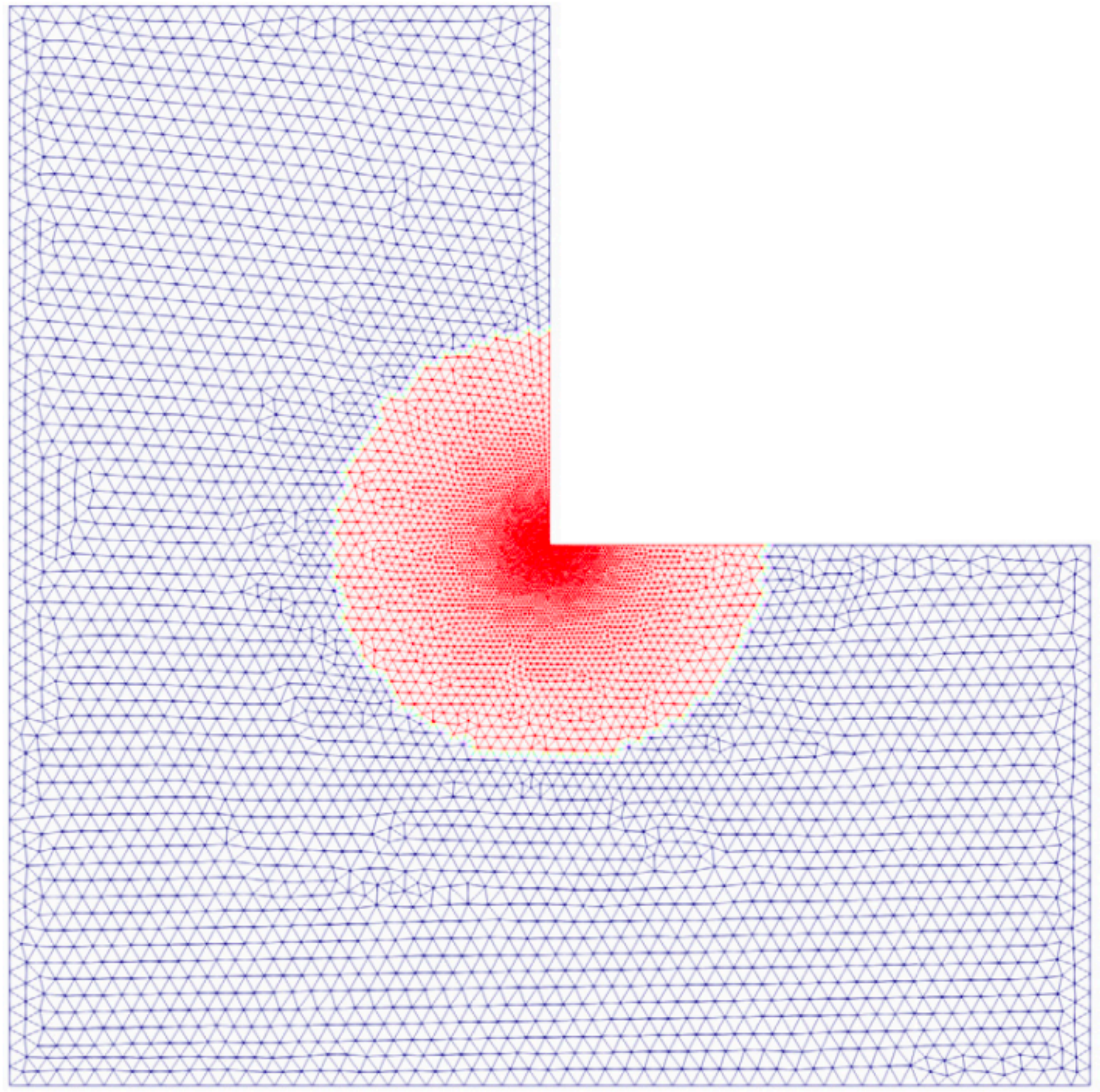
With fixed $s = m = 10$.

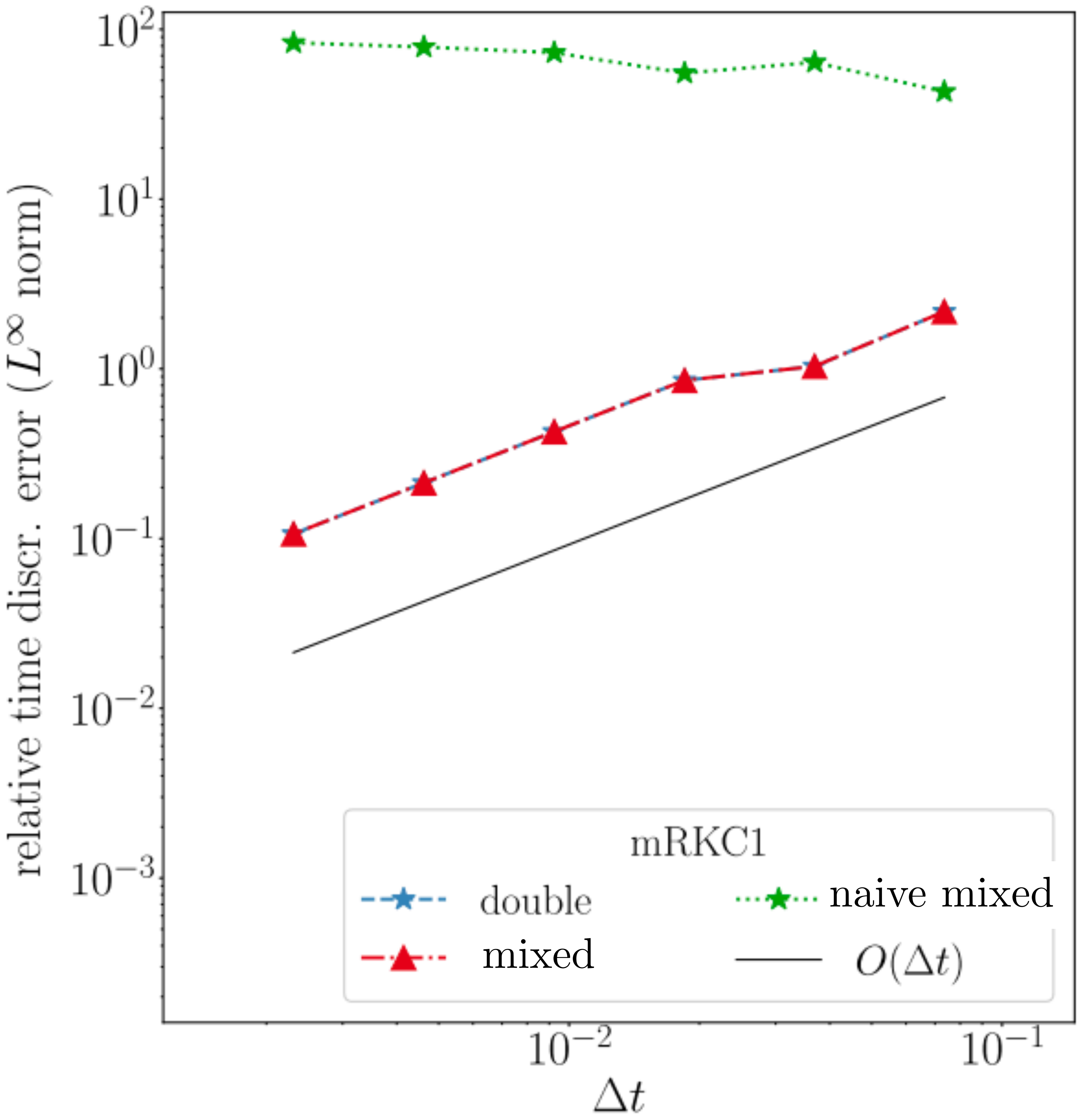


Numerical experiments

Solve

$$\begin{aligned} \frac{\partial u}{\partial t} &= \Delta_F u + \Delta_S u + f_S(x) && \text{in } \Omega \times [0,T], \\ u(x,t) &= 0 && \text{on } \partial\Omega \times [0,T], \\ u(x,0) &= u_0(x) && \text{in } \Omega, \end{aligned}$$





Conclusions

Mixed-precision explicit stabilized methods for

$$y' = f(y)$$

- Only 1 high-precision evaluation of f,
- $s - 1$ evaluations of f in low-precision, with $s = \mathcal{O}(\sqrt{\rho})$.
- Order 1 and 2 methods,
- Order of convergence is preserved (proved),
- Numerically stable.

Conclusions

Multirate mixed-precision explicit stabilized methods for

$$y' = f_F(y) + f_S(y)$$

- Only 1 high-precision evaluation of f_F, f_S,
- $s-1$ evaluations of f_S in low-precision, with s depending on stiffness of f_S only: $s = \mathcal{O}(\sqrt{\rho_S})$.
- $s \cdot m - 1$ evaluations of f_F in low-precision, with $s \cdot m = \mathcal{O}(\sqrt{\rho_F})$
- Order 1 method,
- Order of convergence is preserved (proved),
- Numerically stable.

Bibliography

Thank you!

- Croci, M., & Rosilho de Souza, G. Mixed-precision explicit stabilized Runge-Kutta methods for single- and multi-scale differential equations. *Journal of Computational Physics*, *464*, 2022.

Funding: This project has received funding from the Swiss National Science Foundation, under grant No. 200020_172710 and the European High-Performance Computing Joint Undertaking (JU) under grant agreement No 955701 (TIME-X). The JU receives support from the European Union's Horizon 2020 research and innovation programme and Belgium, France, Germany, and Switzerland.