

# Quadratic functional preserving schemes for linear equations

Hendrik Ranocha<sup>\*1</sup> and Jochen Schütz<sup>†2</sup>

<sup>1</sup>Institute of Mathematics, Johannes Gutenberg University Mainz, Staudingerweg 9, 55128 Mainz, Germany

<sup>2</sup>Faculty of Sciences & Data Science Institute, Hasselt University, Agoralaan Gebouw D, BE-3590 Diepenbeek, Belgium

November 30, 2023

**Key words.**

**AMS subject classification.**

In this work, we consider the *linear* differential equation

$$u'(t) = Au(t) \tag{0.1}$$

for some matrix  $A \in \mathbb{R}^{n \times n}$ ; equipped with suitable initial conditions  $u(t \equiv 0) = u_0 \in \mathbb{R}^n$ . We assume that the quadratic entropy

$$\eta(u) := u^T u \tag{0.2}$$

is conserved, i.e., there holds  $\eta(u(t)) = \eta(u(0))$  for all times  $t$ . Hence, we can compute

$$0 \equiv \frac{d}{dt} \eta(u(t)) = \eta'(u)Au = u^T Au.$$

From this, it is easy to conclude that the matrix  $A$  is skew-symmetric, i.e., there holds

$$A^T = -A. \tag{0.3}$$

Recently in [2], we have observed that a certain class of methods behave very favorably in the context of quadratic entropies, also for nonlinear equations. In this work, we show that for linear equations, these schemes preserve the quadratic entropy explicitly, which obviously makes them highly suited as well for nonlinear equations.

First, in the notation of [2], we consider the class of two-point collocation schemes with  $m$  derivatives on both sides, given by [1, p. II.13]

$$\sum_{k=0}^{2m} \Delta t^k P^{(2m-k)}(0) g^{(k)}(u^{n+1}) = \sum_{k=0}^{2m} \Delta t^k P^{(2m-k)}(1) g^{(k)}(u^n), \tag{0.4}$$

---

<sup>\*</sup>ORCID: 0000-0002-3456-2277

<sup>†</sup>ORCID: 0000-0002-6355-9130

$m$	scheme	
1	$u^{n+1} - \frac{\Delta t}{2} g^{(1)}(u^{n+1})$	$= u^n + \frac{\Delta t}{2} g^{(1)}(u^n)$
2	$u^{n+1} - \frac{\Delta t}{2} g^{(1)}(u^{n+1}) + \frac{\Delta t^2}{12} g^{(2)}(u^{n+1})$	$= u^n + \frac{\Delta t}{2} g^{(1)}(u^n) + \frac{\Delta t^2}{12} g^{(2)}(u^n)$
3	$u^{n+1} - \frac{\Delta t}{2} g^{(1)}(u^{n+1}) + \frac{\Delta t^2}{10} g^{(2)}(u^{n+1}) - \frac{\Delta t^3}{120} g^{(3)}(u^{n+1})$	$= u^n + \frac{\Delta t}{2} g^{(1)}(u^n) + \frac{\Delta t^2}{10} g^{(2)}(u^n) + \frac{\Delta t^3}{120} g^{(3)}(u^n)$

Table 1: The first three schemes of form (0.4).

where  $P$  can be taken as the polynomial

$$P(t) = \frac{t^m(t-1)^m}{(2m)!}.$$

In here, we have defined  $g^{(0)}(u) := u$ ,  $g^{(1)}(u) := Au$ , and  $g^{(k)}(u)$  as the  $k$ -th temporal derivative of an exact solution  $u$  to (0.1). For all  $k$ , this amounts to

$$g^{(k)}(u) := A^k u, \quad \forall k \in \mathbb{N}^{\geq 0}. \quad (0.5)$$

In Tbl. 1, the first few schemes have been listed.

For the analysis to follow, we need the following Lemma. Its proof is an obvious consequence of the fact that the function  $P$  is symmetric w.r.t. the point  $t = \frac{1}{2}$ .

**Lemma 0.1.** *Let  $0 \leq k \leq 2m$ . There holds  $P^{(2m-k)}(1) = P^{(2m-k)}(0)$  for an even  $k$ , and  $P^{(2m-k)}(1) = -P^{(2m-k)}(0)$  for an uneven  $k$ .*

For all these schemes,  $\eta$  from Eq. (0.2) is conserved if there holds  $A^T = -A$ , i.e., there holds  $\eta(u^{n+1}) = \eta(u^n)$  if  $u^{n+1}$  has been computed according to (0.4):

**Theorem 0.2.** *Given that  $u^{n+1}$  is computed according to (0.4) and there holds Eq. (0.3), there holds*

$$\eta(u^{n+1}) = \eta(u^n) \quad (0.6)$$

for a quadratic functional of form (0.2).

*Proof.* Define  $\alpha_k := P^{(2m-k)}(1)$  and  $\alpha_k^1 := P^{(2m-k)}(0)$  for  $0 \leq k \leq m$  and note that  $P^{(2m-k)}(1) = P^{(2m-k)}(0) = 0$  for  $k > m$ . Furthermore, due to La. 0.1, there holds  $\alpha_k = (-1)^k \alpha_k^1$ . Hence, (0.4) can be written as

$$\sum_{k=0}^m (-1)^k \Delta t^k \alpha_k g^{(k)}(u^{n+1}) = \sum_{k=0}^m \Delta t^k \alpha_k g^{(k)}(u^n). \quad (0.7)$$

Using (0.5), this leads to the linear equation

$$\underbrace{\left( \sum_{k=0}^m (-1)^k \Delta t^k \alpha_k A^k \right)}_{=: \mathcal{A}} u^{n+1} = \underbrace{\left( \sum_{k=0}^m \Delta t^k \alpha_k A^k \right)}_{=: \mathcal{B}} u^n. \quad (0.8)$$

From this, we can compute the entropy as

$$\eta(u^{n+1}) = (u^{n+1})^T u^{n+1} = (\mathcal{A}^{-1} \mathcal{B} u^n)^T \mathcal{A}^{-1} \mathcal{B} u^n = (u^n)^T \mathcal{B}^T \mathcal{A}^{-T} \mathcal{A}^{-1} \mathcal{B} u^n.$$

It hence remains to show that  $(\mathcal{A}^{-1} \mathcal{B})^T \mathcal{A}^{-1} \mathcal{B}$  is the identity, which is equivalent to  $\mathcal{A} \mathcal{A}^T = \mathcal{B} \mathcal{B}^T$ . Exploiting the property (0.3), from which there follows that

$$(A^k)^T = (-1)^k A^k,$$

we obtain

$$\begin{aligned}\mathcal{A}^T &= \sum_{k=0}^m (-1)^k \Delta t^k \alpha_k (A^k)^T = \sum_{k=0}^m \Delta t^k \alpha_k A^k = \mathcal{B}, \\ \mathcal{B}^T &= \sum_{k=0}^m \Delta t^k \alpha_k (A^k)^T = \sum_{k=0}^m (-1)^k \Delta t^k \alpha_k A^k = \mathcal{A}.\end{aligned}\tag{0.9}$$

Hence, there holds

$$\mathcal{A}\mathcal{A}^T = \mathcal{A}\mathcal{B} = \mathcal{B}^T\mathcal{B}.$$

$\mathcal{B}^T$  and  $\mathcal{B}$  are commuting matrices, as they both can be written as matrix polynomials of the same matrix  $A$ , see (0.9). Hence, the result follows.  $\square$

We consider the scheme HB-I2DRK6-3s, the sixth-order two-stage two-derivative Runge-Kutta scheme. The first stage is explicit, the third stage is equal to  $u^{n+1}$ , so only one intermediate stage remains that we denote by  $\bar{u}$ . Hence, the scheme is given by the implicit equation

$$\begin{aligned}\bar{u} &= \left( \text{Id} + \Delta t B_{21}^{(1)} A + \Delta t^2 B_{21}^{(2)} A^2 \right) u^n + \left( \Delta t B_{22}^{(1)} A + \Delta t^2 B_{22}^{(2)} A^2 \right) \bar{u} + \left( \Delta t B_{23}^{(1)} A + \Delta t^2 B_{23}^{(2)} A^2 \right) u^{n+1} \\ u^{n+1} &= \left( \text{Id} + \Delta t B_{31}^{(1)} A + \Delta t^2 B_{31}^{(2)} A^2 \right) u^n + \left( \Delta t B_{32}^{(1)} A + \Delta t^2 B_{32}^{(2)} A^2 \right) \bar{u} + \left( \Delta t B_{33}^{(1)} A + \Delta t^2 B_{33}^{(2)} A^2 \right) u^{n+1}\end{aligned}$$

with Butcher tableaux given by

$$B^{(1)} = \begin{pmatrix} 0 & 0 & 0 \\ \frac{101}{480} & \frac{8}{30} & \frac{55}{2400} \\ \frac{7}{30} & \frac{16}{30} & \frac{7}{30} \end{pmatrix}, \quad B^{(2)} = \begin{pmatrix} 0 & 0 & 0 \\ \frac{65}{4800} & -\frac{25}{600} & -\frac{25}{8000} \\ \frac{5}{300} & 0 & -\frac{5}{300} \end{pmatrix}. \quad (0.10)$$

For compactness, define

$$\mathcal{B}_{ik} := \Delta t B_{ik}^{(1)} A + \Delta t^2 B_{ik}^{(2)} A^2.$$

Then, the linear system of equations to be solved is given by

$$\begin{pmatrix} \text{Id} - \mathcal{B}_{22} & -\mathcal{B}_{23} \\ -\mathcal{B}_{32} & \text{Id} - \mathcal{B}_{33} \end{pmatrix} \begin{pmatrix} \bar{u} \\ u^{n+1} \end{pmatrix} = \begin{pmatrix} (\text{Id} + \mathcal{B}_{21}) u^n \\ (\text{Id} + \mathcal{B}_{31}) u^n \end{pmatrix}$$

First, we eliminate the term  $\bar{u}$ ; it can be written in terms of  $u^n$  and  $u^{n+1}$  as

$$\bar{u} = (\text{Id} - \mathcal{B}_{22})^{-1} \left( (\text{Id} + \mathcal{B}_{21}) u^n + \mathcal{B}_{23} u^{n+1} \right).$$

Then, the remaining equation in  $u^{n+1}$  is given by

$$\underbrace{\left( -\mathcal{B}_{32} (\text{Id} - \mathcal{B}_{22})^{-1} \mathcal{B}_{23} + \text{Id} - \mathcal{B}_{33} \right)}_{:=\mathcal{A}} u^{n+1} = \underbrace{\left( \text{Id} + \mathcal{B}_{31} + \mathcal{B}_{32} (\text{Id} - \mathcal{B}_{22})^{-1} (\text{Id} + \mathcal{B}_{21}) \right)}_{:=\mathcal{B}} u^n.$$

Now, in the same way as above, we can see that

$$\eta(u^{n+1}) = (u^n)^T \mathcal{B}^T \mathcal{A}^{-T} \mathcal{A}^{-1} \mathcal{B} u^n,$$

and we have to prove that  $\mathcal{A} \mathcal{A}^T = \mathcal{B} \mathcal{B}^T$  as before.

**Lemma 0.3.** *The following is true for each matrix  $A$  given that (0.3) holds:*

1.  $\mathcal{B}_{31}^T = -\mathcal{B}_{33}$ ,
2.  $\mathcal{B}_{32}^T = -\mathcal{B}_{32}$ ,
3.  $\mathcal{S} - \mathcal{S}^T = \mathcal{B}_{32} \mathcal{S} \mathcal{S}^T$ .

5

5

5

5

5

5

5

5

5

5

5

Factoring out

$$\begin{aligned} & \mathcal{B}_{32}\mathcal{S}\mathcal{B}_{23}\mathcal{B}_{23}^T\mathcal{S}^T\mathcal{B}_{32}^T - \mathcal{B}_{32}\mathcal{S}\mathcal{B}_{23} + \mathcal{B}_{32}\mathcal{S}\mathcal{B}_{23}\mathcal{B}_{33}^T - \mathcal{B}_{23}^T\mathcal{S}^T\mathcal{B}_{32}^T + \mathcal{B}_{33}\mathcal{B}_{23}^T\mathcal{S}^T\mathcal{B}_{32}^T \stackrel{!}{=} \\ & \mathcal{S}^T\mathcal{B}_{32}^T + \mathcal{B}_{21}^T\mathcal{S}^T\mathcal{B}_{32}^T - \mathcal{B}_{33}^T\mathcal{S}^T\mathcal{B}_{32}^T - \mathcal{B}_{33}^T\mathcal{B}_{21}^T\mathcal{S}^T\mathcal{B}_{32}^T + \mathcal{B}_{32}\mathcal{S} + \mathcal{B}_{32}\mathcal{S}\mathcal{B}_{21} - \mathcal{B}_{32}\mathcal{S}\mathcal{B}_{33} - \mathcal{B}_{32}\mathcal{S}\mathcal{B}_{21}\mathcal{B}_{33} \\ & + \mathcal{B}_{32}\mathcal{S}\mathcal{S}^T\mathcal{B}_{32}^T + \mathcal{B}_{32}\mathcal{S}\mathcal{B}_{21}\mathcal{S}^T\mathcal{B}_{32}^T + \mathcal{B}_{32}\mathcal{S}\mathcal{B}_{21}^T\mathcal{S}^T\mathcal{B}_{32}^T + \mathcal{B}_{32}\mathcal{S}\mathcal{B}_{21}\mathcal{B}_{21}^T\mathcal{S}^T\mathcal{B}_{32}^T \end{aligned}$$

Now use the fact that  $\mathcal{S} - \mathcal{S}^T = \mathcal{B}_{32}\mathcal{S}\mathcal{S}^T$  and  $\mathcal{B}_{32}^T = -\mathcal{B}_{32}$ :

$$\begin{aligned} & \mathcal{B}_{32}\mathcal{S}\mathcal{B}_{23}\mathcal{B}_{23}^T\mathcal{S}^T\mathcal{B}_{32}^T - \mathcal{B}_{32}\mathcal{S}\mathcal{B}_{23} + \mathcal{B}_{32}\mathcal{S}\mathcal{B}_{23}\mathcal{B}_{33}^T - \mathcal{B}_{23}^T\mathcal{S}^T\mathcal{B}_{32}^T + \mathcal{B}_{33}\mathcal{B}_{23}^T\mathcal{S}^T\mathcal{B}_{32}^T \stackrel{!}{=} \\ & \mathcal{B}_{21}^T\mathcal{S}^T\mathcal{B}_{32}^T - \mathcal{B}_{33}^T\mathcal{S}^T\mathcal{B}_{32}^T - \mathcal{B}_{33}^T\mathcal{B}_{21}^T\mathcal{S}^T\mathcal{B}_{32}^T + \mathcal{B}_{32}\mathcal{S}\mathcal{B}_{21} - \mathcal{B}_{32}\mathcal{S}\mathcal{B}_{33} - \mathcal{B}_{32}\mathcal{S}\mathcal{B}_{21}\mathcal{B}_{33} \\ & + \mathcal{B}_{32}\mathcal{S}\mathcal{B}_{21}\mathcal{S}^T\mathcal{B}_{32}^T + \mathcal{B}_{32}\mathcal{S}\mathcal{B}_{21}^T\mathcal{S}^T\mathcal{B}_{32}^T + \mathcal{B}_{32}\mathcal{S}\mathcal{B}_{21}\mathcal{B}_{21}^T\mathcal{S}^T\mathcal{B}_{32}^T \end{aligned}$$

In a similar way, we can make use of the fact that  $\mathcal{S}\mathcal{B}_{23} - \mathcal{B}_{21}^T\mathcal{S}^T - \mathcal{S}\mathcal{B}_{33} - \mathcal{S}\mathcal{B}_{21}^T\mathcal{S}^T\mathcal{B}_{32} = 0$  **TODO: I did the proof, fully analogously to the one with  $\mathcal{S} - \mathcal{S}^T = \mathcal{B}_{32}\mathcal{S}\mathcal{S}^T$  .!**, resulting in

$$\begin{aligned} & \mathcal{B}_{32}\mathcal{S}\mathcal{B}_{23}\mathcal{B}_{23}^T\mathcal{S}^T\mathcal{B}_{32}^T + \mathcal{B}_{32}\mathcal{S}\mathcal{B}_{23}\mathcal{B}_{33}^T - \mathcal{B}_{23}^T\mathcal{S}^T\mathcal{B}_{32}^T + \mathcal{B}_{33}\mathcal{B}_{23}^T\mathcal{S}^T\mathcal{B}_{32}^T \stackrel{!}{=} \\ & -\mathcal{B}_{33}^T\mathcal{S}^T\mathcal{B}_{32}^T - \mathcal{B}_{33}^T\mathcal{B}_{21}^T\mathcal{S}^T\mathcal{B}_{32}^T + \mathcal{B}_{32}\mathcal{S}\mathcal{B}_{21} - \mathcal{B}_{32}\mathcal{S}\mathcal{B}_{21}\mathcal{B}_{33} + \mathcal{B}_{32}\mathcal{S}\mathcal{B}_{21}\mathcal{S}^T\mathcal{B}_{32}^T + \mathcal{B}_{32}\mathcal{S}\mathcal{B}_{21}\mathcal{B}_{21}^T\mathcal{S}^T\mathcal{B}_{32}^T \end{aligned}$$

Now, we use  $\mathcal{B}_{23}^T\mathcal{S}^T + \mathcal{B}_{33}^T\mathcal{S}^T + \mathcal{S}\mathcal{B}_{21} - \mathcal{S}\mathcal{B}_{21}\mathcal{S}^T\mathcal{B}_{32} = 0$  **TODO: proved, also similar!:**

$$\mathcal{B}_{32}\mathcal{S}\mathcal{B}_{23}\mathcal{B}_{23}^T\mathcal{S}^T\mathcal{B}_{32}^T + \mathcal{B}_{32}\mathcal{S}\mathcal{B}_{23}\mathcal{B}_{33}^T + \mathcal{B}_{33}\mathcal{B}_{23}^T\mathcal{S}^T\mathcal{B}_{32}^T \stackrel{!}{=} -\mathcal{B}_{33}^T\mathcal{B}_{21}^T\mathcal{S}^T\mathcal{B}_{32}^T - \mathcal{B}_{32}\mathcal{S}\mathcal{B}_{21}\mathcal{B}_{33} + \mathcal{B}_{32}\mathcal{S}\mathcal{B}_{21}\mathcal{B}_{21}^T\mathcal{S}^T\mathcal{B}_{32}^T$$

Also this term is equal, which can be seen with similar arguments as before.

One of the things that experimentally seems to hold true is that

$$\mathcal{S} - \mathcal{S}^T = \mathcal{B}_{32} \mathcal{S} \mathcal{S}^T.$$

I try to prove this here. **TODO: The proof to follow is prototypical for all the other things that we show!**

$$\mathcal{S} \mathcal{S}^T = \left( \left( \text{Id} - \frac{4\Delta t}{15} A + \frac{\Delta t^2}{24} A^2 \right) \left( \text{Id} + \frac{4\Delta t}{15} A + \frac{\Delta t^2}{24} A^2 \right) \right)^{-1}$$

We multiply out the interior:

$$\begin{aligned} & \left( \text{Id} - \frac{4\Delta t}{15} A + \frac{\Delta t^2}{24} A^2 \right) \left( \text{Id} + \frac{4\Delta t}{15} A + \frac{\Delta t^2}{24} A^2 \right) \\ &= \text{Id} + \frac{4\Delta t}{15} A + \frac{\Delta t^2}{24} A^2 - \frac{4\Delta t}{15} A - \frac{16\Delta t^2}{15 \cdot 15} A^2 - \frac{4\Delta t^3}{15 \cdot 24} A^3 + \frac{\Delta t^2}{24} A^2 + \frac{4\Delta t^3}{24 \cdot 15} A^3 + \frac{\Delta t^4}{24^2} A^4 \\ &= \text{Id} + \frac{\Delta t^2}{24} A^2 - \frac{16\Delta t^2}{15 \cdot 15} A^2 + \frac{\Delta t^2}{24} A^2 + \frac{\Delta t^4}{24^2} A^4 \\ &= \text{Id} + \frac{11\Delta t^2}{900} A^2 + \frac{\Delta t^4}{24^2} A^4, \end{aligned}$$

which turns out to be an even, i.e., symmetric, quantity. Now, there holds

$$\mathcal{B}_{32} \mathcal{S} \mathcal{S}^T = \frac{8\Delta t}{15} A \left( \text{Id} + \frac{11\Delta t^2}{900} A^2 + \frac{\Delta t^4}{24^2} A^4 \right)^{-1}$$

and

$$\mathcal{S} - \mathcal{S}^T = \left( \text{Id} - \frac{4\Delta t}{15} A + \frac{\Delta t^2}{24} A^2 \right)^{-1} - \left( \text{Id} + \frac{4\Delta t}{15} A + \frac{\Delta t^2}{24} A^2 \right)^{-1}.$$

$\mathcal{B}_{33} \mathcal{S} \mathcal{S}^T = \mathcal{S} - \mathcal{S}^T$  is now equivalent to (note that all the matrices commute!)

$$\frac{8\Delta t}{15} A \left( \text{Id} - \frac{4\Delta t}{15} A + \frac{\Delta t^2}{24} A^2 \right) \left( \text{Id} + \frac{4\Delta t}{15} A + \frac{\Delta t^2}{24} A^2 \right) = \left( \text{Id} + \frac{11\Delta t^2}{900} A^2 + \frac{\Delta t^4}{24^2} A^4 \right) \left( \text{Id} + \frac{4\Delta t}{15} A + \frac{\Delta t^2}{24} A^2 \right) - \left( \text{Id} + \frac{11\Delta t^2}{900} A^2 + \frac{\Delta t^4}{24^2} A^4 \right) \left( \text{Id} - \frac{4\Delta t}{15} A + \frac{\Delta t^2}{24} A^2 \right)$$

It is an easy, yet tedious, task to confirm that this is equal.

## References

- [1] E. Hairer, S. P. Nørsett, and G. Wanner. *Solving Ordinary Differential Equations I: Nonstiff Problems*. Vol. 8. Springer Series in Computational Mathematics. Berlin Heidelberg: Springer-Verlag, 2008. doi: 10.1007/978-3-540-78862-1.
- [2] H. Ranocha and J. Schütz. *Multiderivative time integration methods preserving nonlinear functionals via relaxation*. 2023. arXiv: 2311.03883 [math.NA].