from Applied Numer. Math. **17**, 293-298 (1995).

MONOTONICITY OF QUADRATIC FORMS WITH SYMPLECTIC RUNGE–KUTTA METHODS.

TIMO EIROLA

1. Introduction

It is well known that the Gauss (Gauss–Legendre) methods are algebraically stable, and since this implies B–stability they produce contractive numerical maps when applied to contractive systems ([5]). Also it is known that they (or more generally symplectic RK's) preserve any quadratic integrals of the system . In particular, they produce symplectic maps for Hamiltonian systems ([6]). Further, in [2] it was noticed that they can be used to preserve positive definiteness of the numerical solution of Riccati differential equations in the Hamiltonian form.

In [7] Humphries and Stuart noticed connections between some of these results and in [1] Bochev and Scovel showed that symplecticity is a consequence of preservation of quadratic forms and the commutativity of discretation and differentiation with respect to initial condition. Here we demonstrate that all these results follow from a general monotonicity property of these methods for quadratic forms.

2. The general result

Consider an autonomous differential equation

$$\dot{x}(t) = f(x(t)) \ , \qquad x(t) \in \Omega \subset \mathbf{R}^n \ , \tag{2.1}$$

where $f \in C^k(\Omega)$, Ω open, and $k \geq 1$. Assume that the solutions can be extended to the whole time axis. For simplicity we assume that a nonautonomous equation is written in this form with one component of f being identically 1. Denote by ϕ_t , $t \in \mathbf{R}$ the flow of f , i.e.,

$$\partial_t \phi_t(x) = f(\phi_t(x)) \qquad , \quad \phi_0(x) = x \ .$$

Consider an s–stage Runge–Kutta method, applied to (2.1):

$$\psi_h(x) = x + h \sum_{i=1}^{s} b_i f(p_i) \quad , \qquad p_i = x + h \sum_{j=1}^{s} a_{ij} f(p_j) \ . \tag{2.2}$$

We say that the RK–step is *well defined* for $h > 0$, $x \in \Omega$, if the Runge–Kutta stages p_i (and consequently $\psi_h(x)$) are uniquely defined by the implicit function theorem via continuation from $h = 0$, $p_i = x$.

We will call ψ_h the *numerical flow* , even though it is not a flow.

Remark 2.1. For given x let $\bar{h}(x)$ be the supremum of those h such that the RK–step is well defined for h, x. Then it is easy to show that $\bar{h}$ is lower semicontinuous and it follows that $\bar{h}(C)$ is positive for any compact $C \subset \Omega$. If $p \leq k$ is the order of the method, then for $0 < h \leq h_0 < \bar{h}(C)$ we have that ψ_h is $O(h^p)$ C^{k-p}–close to ϕ_h in C.

We assume that the method is *symplectic (canonical)*, i.e., the coefficients satisfy:

$$m_{ij} := b_i a_{ij} + b_j a_{ji} - b_i b_j = 0 \quad \forall\, i, j = 1, \ldots, s\ , \tag{2.3}$$

and that the b_i's are all nonnegative. The Gauss methods are such. Note that for algebraic stability the requirement on the matrix $\{m_{ij}\}$ is only that it is positive semidefinite.

Let $\mathcal{E}$ be a space of functions from a set S to $\mathbf{R}^n$ and $\mathcal{E}_\Omega$ the subset of functions which have compact ranges in Ω. Assume $\mathcal{E}$ has a topology such that ϕ_t, when considered as a flow in $\mathcal{E}_\Omega$:

$$\eta \in \mathcal{E}_\Omega \longrightarrow \phi_t \circ \eta \in \mathcal{E}_\Omega \quad \text{is continuous and} \quad \lim_{t\to 0} \frac{1}{t}[\phi_t \circ \eta - \eta] = f \circ \eta \tag{2.4}$$

for every $\eta \in \mathcal{E}_\Omega$.

Remark 2.2. We defined the functions of $\mathcal{E}_\Omega$ to have compact ranges in order to have positive h for every $\eta \in \mathcal{E}_\Omega$.

Let $\mathcal{Y}$ be a topological vector space with a closed convex cone K (i.e., $a\bar{K} + b\bar{K} \subset K$ for every $a, b > 0$), which we will call positive and for $y, \tilde{y} \in \mathcal{Y}$ denote $y \succeq \tilde{y}$ if $y - \tilde{y} \in K$. In some of the applications K will be the trivial cone containing just the origin.

Let now β be a continuous bilinear map from $\mathcal{E} \times \mathcal{E}$ to $\mathcal{Y}$, and $q(\eta) := \beta(\eta, \eta)$ the corresponding quadratic map.

We take this somewhat abstract setting in order to get many results as corollaries of one theorem.

Theorem 2.1. *Assume that q is nondecreasing along the trajectories of ϕ_t in $\mathcal{E}_\Omega$, i.e.,*

$$q(\phi_t \circ \eta) \succeq q(\eta)\ .$$

for any $t > 0$ and $\eta \in \mathcal{E}_\Omega$. Then also

$$q(\psi_h \circ \eta) \succeq q(\eta)$$

for every $\eta \in \mathcal{E}_\Omega$ and $h > 0$ for which the RK–step is well defined.

Proof. By continuity of β and by (2.4) we have for any $\eta \in \mathcal{E}_\Omega$

$$\begin{aligned} 0 \preceq \lim_{t\to 0} \frac{1}{t}[q(\phi_t \circ \eta) - q(\eta)] &= \lim_{t\to 0} \Big[\beta\big(\frac{\phi_t \circ \eta - \eta}{t}, \phi_t \circ \eta\big) + \beta\big(\eta\ ,\ \frac{\phi_t \circ \eta - \eta}{t}\big)\Big] \\ &= \beta(f \circ \eta\ ,\ \eta) + \beta(\eta, f \circ \eta)\ . \end{aligned} \tag{2.5}$$

The Runge–Kutta method produces:

$$\psi_h \circ \eta = \eta + h \sum_{i=1}^{s} b_i f \circ \pi_i \quad , \qquad \pi_i = \eta + h \sum_{j=1}^{s} a_{ij} f \circ \pi_j \quad . \tag{2.6}$$

Thus we get:

$$(2.7)\quad q(\psi_h \circ \eta) = \beta\Big(\eta + h\sum_{i=1}^{s} b_i f\circ\pi_i\ ,\ \eta + h\sum_{j=1}^{s} b_j f\circ\pi_j\Big) =$$

$$= q(\eta) + h\Big[\sum_{i=1}^{s} b_i\beta(f\circ\pi_i,\eta) + \sum_{j=1}^{s} b_j\beta(\eta, f\circ\pi_j)\Big] + h^2\sum_{i,j=1}^{s} b_i b_j \beta(f\circ\pi_i, f\circ\pi_j)\ .$$

From the latter of equations (2.6) we get

$$\beta(f\circ\pi_i,\eta) = \beta(f\circ\pi_i,\pi_i) - h\sum_{j=1}^{s} a_{ij}\beta(f\circ\pi_i, f\circ\pi_j)\,,$$

$$\beta(\eta, f\circ\pi_j) = \beta(\pi_j, f\circ\pi_j) - h\sum_{i=1}^{s} a_{ji}\beta(f\circ\pi_i, f\circ\pi_j)\,.$$

Putting these in (2.7) gives

$$(2.8)\quad q(\psi_h\circ\eta) = q(\eta) + h\sum_{i=1}^{s} b_i\left[\beta(f\circ\pi_i,\pi_i) + \beta(\pi_i, f\circ\pi_i)\right] - h^2\sum_{i,j=1}^{s} m_{ij}\,\beta(f\circ\pi_i, f\circ\pi_j)\,.$$

Now, by (2.3) the h^2 –terms vanish and the assertion follows from (2.5) and positivity of the b_i's. □

Remark 2.3. There is nothing really new in the proof above. The essential algebraic formula (2.8) from which the result can be read (and which holds for any RK–method) can be seen in the proofs of all the results mentioned in the beginning.

3. Applications

In the following applications we either have S a singleton set, when $\mathcal{E}$ can be identified with $\mathbf{R}^n$, or $\mathcal{E} = C^r(\Omega, \mathbf{R}^n)$, $0 \le r \le k$, with the topology of uniform converge on compact sets of derivatives up to order r. It is easy to show that (2.4) holds in these cases.

Remark 3.1. For B-stability, i.e., q being a negative semidefinite quadratic form in $\mathbf{R}^n$ such that $q(\phi_t(x) - \phi_t(x'))$ is nondecreasing for any two initial points x, x', taking $\mathcal{E} = R^{2n}$ one sees from (2.8) (since the Kronecker product of two symmetric semidefinite matrices is semidefinite) that it suffices for $\{m_{ij}\}$ to be positive semidefinite.

3.1. Conservation of quadratic integrals. Assume that the flow satisfies

$$\langle\phi_t(x), Q\phi_t(x)\rangle = \langle x, Qx\rangle \quad \text{, i.e.,} \qquad x^T(Q+Q^T)f(x) = 0$$

for every t, x, where Q is a constant matrix. Take $\mathcal{E} = \mathbf{R}^n$, $\beta(x,y) = \langle x, Qy\rangle$, and set $\mathcal{Y} = \mathbf{R}$, $K = \{0\}$. Then the theorem says that

$$\langle\psi_h(x), Q\psi_h(x)\rangle = \langle x, Qx\rangle$$

for every h, x for which the RK–step is well defined, i.e., the quadratic form is preserved also by the numerical flow.

An application of this is that when one solves $\dot{U}(t) = S(U(t))U(t)$ with these methods for an $n \times n$ –matrix U in the situation, where $S(U)$ is skew–symmetric for every U, then the numerical solution will (as well as the true one) always be an orthogonal matrix, provided the initial condition is (see [4]).

3.2. Symplecticity. Assume that f is Hamiltonian, i.e., n is even and

$$f = J\nabla H \quad , \text{ where } \quad J = \begin{bmatrix} 0 & I \\ -I & 0 \end{bmatrix} .$$

It follows that ϕ_t is a *symplectic map*:

$$(\partial_x \phi_t(x))^T J \partial_x \phi_t(x) = J$$

for every t, x. Take $\mathcal{E} = C^1(\Omega, \mathbf{R}^n)$ and $\beta(\xi, \eta) = \xi_x(\cdot)^T J \eta_x(\cdot)$, and set $\mathcal{Y} = C(\Omega, \mathbf{R}^{n\times n})$, $K = \{0\}$. Then the theorem says that for every η in $\mathcal{E}_\Omega$ we have

$$\eta_x^T (\partial_x \psi_h)^T J (\partial_x \psi_h) \eta_x = \eta_x^T J \eta_x .$$

Thus $(\partial_x \psi_h(x))^T J \partial_x \psi_h(x) = J$, i.e., ψ_h is also a symplectic map.

3.3. Positive definiteness for Riccati equations. Consider solving numerically the symmetric matrix Riccati equation

$$\dot{X}(t) = A(t)X(t) + X(t)A(t)^T - X(t)B(t)X(t) + C(t) , \tag{3.1}$$

in $\mathbf{R}^{d\times d}$ with a symmetric positive semidefinite initial condition $X(0) = X_0$. We assume that the matrices B and C are symmetric and positive semidefinite. Then the solution of (3.1) is symmetric and positive semidefinite. It was shown in [2] that essentially only the backward Euler method preserves this property when applied to (3.1).

On the other hand it is well known that the solution of (3.1) can also be obtained as $X(t) = Y(t)Z(t)^{-1}$, where

$$\dot{\begin{bmatrix} Y(t) \\ Z(t) \end{bmatrix}} = \begin{bmatrix} A(t) & C(t) \\ B(t) & -A(t)^T \end{bmatrix} \begin{bmatrix} Y(t) \\ Z(t) \end{bmatrix} \quad , \quad \begin{bmatrix} Y(0) \\ Z(0) \end{bmatrix} = \begin{bmatrix} X_0 \\ I \end{bmatrix} . \tag{3.2}$$

Denote the matrix above by M. To show the nice properties of canonical methods for this equation take $\mathcal{E} = \mathbf{R}^{2d\times d}$, $\mathcal{Y} = \mathbf{R}^{d\times d}$, let K be the cone of positive semidefinite $d \times d$ –matrices, and set

$$\beta\left(\begin{bmatrix} Y \\ Z \end{bmatrix}, \begin{bmatrix} \tilde{Y} \\ \tilde{Z} \end{bmatrix}\right) := Z^T \tilde{Y}.$$

Then

$$\beta\left(M \begin{bmatrix} Y \\ Z \end{bmatrix}, \begin{bmatrix} Y \\ Z \end{bmatrix}\right) + \beta\left(\begin{bmatrix} Y \\ Z \end{bmatrix}, M \begin{bmatrix} Y \\ Z \end{bmatrix}\right) = (BY - A^T Z)^T Y + Z^T(AY + CZ) = Y^T BY + Z^T CZ \succeq 0 ,$$

i.e., q is increasing along the trajectories. This is then true also for the numerical trajectories (Y_k, Z_k). Hence $Z_k^T Y_k$ and consequently $X_k := Y_k(Z_k)^{-1} = (Z_k^T)^{-1} Z_k^T Y_k Z_k^{-1}$ will be positive semidefinite (definite if the initial condition is).

Similarly, taking $\hat{K} = 0$ and $\hat{\beta}\left(\left[\begin{smallmatrix} Y \\ Z \end{smallmatrix}\right], \left[\begin{smallmatrix} \tilde{Y} \\ \tilde{Z} \end{smallmatrix}\right]\right) := Z^T \tilde{Y} - Y^T \tilde{Z}$ one easily shows that the numerical solution $X_k := Y_k Z_k^{-1}$ is also symmetric.

Monotonicity: If in (3.1) one perturbs matrices B, C, X_0 (pointwise) to $\tilde{B} \preceq B, \tilde{C} \succeq C, \tilde{X}_0 \succeq X_0$ without changing A, then it is easy to show that the corresponding solution also satisfies $\tilde{X}(t) \succeq X(t)$. The same property for the numerical solution is also true if one obtains it by applying a canonical scheme to (3.2). This too can be shown via a (less simple) application of the previous theorem (see [3]).

Remark 3.2. (Warning) In the beginning of this work the author believed that *conformality* would also be preserved by these methods. This means the property that

$$(\partial_x \phi_t(x))^T \partial_x \phi_t(x) = \alpha(t, x) I$$

for some scalar function α. This holds if the Jacobian of f is of the form: skew–symmetric plus scalar times identity. This, however, is not a property that could be written as monotonicity of a quadratic map (unless $\alpha \equiv 1$) and the Gauss methods do not preserve it.

References

1. P. B. Bochev and C. Scovel, *On quadratic invariants and symplectic structure*, BIT (1994).
2. L. Dieci and T. Eirola, *Positive definiteness in the numerical solution of Riccati differential equations*, Numer. Mathematik **67** (1994), 303–313.
3. ———, *Preserving monotonicity in the numerical solution of Riccati differential equations*, to appear (1995).
4. L. Dieci, R. D. Russell, and E. S. van Vleck, *Unitary integrators and applications to continuous orthonormalization techniques*, SIAM J. Numer. Anal. **31** (1994), 261–281.
5. E. Hairer and G. Wanner, *Solving ordinary differential equations II*, Springer-Verlag, Berlin-Heidelberg, 1991.
6. J. M. Sanz-Serna, *Symplectic integrators for Hamiltonian problems: An overview*, Acta Numerica **1** (1992), 243–286.
7. A. M. Stuart and A. R. Humphries, *Model problems in numerical stability theory for initial value problems*, SIAM Review **36** (1994), 226–257.

Inst. of Mathematics, Helsinki Univ. of Technology, SF-02150 Espoo, Finland

E-mail address: Timo.Eirola@hut.fi