SIAM J. NUMER. ANAL.
Vol. 31, No. 1, pp. 261–281, February 1994

014

UNITARY INTEGRATORS AND APPLICATIONS TO CONTINUOUS ORTHONORMALIZATION TECHNIQUES*

LUCA DIECI†, ROBERT D. RUSSELL‡, AND ERIK S. VAN VLECK‡§

Abstract. In this paper the issue of integrating matrix differential systems whose solutions are unitary matrices is addressed. Such systems have skew-Hermitian coefficient matrices in the linear case and a related structure in the nonlinear case. These skew systems arise in a number of applications, and interest originates from application to continuous orthogonal decoupling techniques. In this case, the matrix system has a cubic nonlinearity.

Numerical integration schemes that compute a unitary approximate solution for all stepsizes are studied. These schemes can be characterized as being of two classes: *automatic* and *projected unitary schemes.* In the former class, there belong those standard finite difference schemes which give a unitary solution; the only ones are in fact the Gauss–Legendre point Runge–Kutta (Gauss RK) schemes. The second class of schemes is created by projecting approximations computed by an arbitrary scheme into the set of unitary matrices. In the analysis of these unitary schemes, the stability considerations are guided by the skew-Hermitian character of the problem. Various error and implementation issues are considered, and the methods are tested on a number of examples.

Key words. unitary integrators, structure preserving algorithms, continuous orthonormalization

AMS subject classification. 65L

Notation. We consider matrices $A \in \mathbb{C}^{n\times p}$, $A = (a_{ij})_{i,j=1}^{n,p}$, $a_{ij} \in \mathbb{C}$. We say that A is Hermitian if $A = A^*$, $A^* = \bar{A}^T$ (the conjugate transpose), and write $A \in \mathcal{H}^{n\times n}$. S is skew-Hermitian if $S^* = -S$, and we write $S \in \mathcal{S}^{n\times n}$. $U \in \mathbb{C}^{n\times p}$ is unitary if $U^*U = I$, and we write $U \in \mathcal{U}^{n\times p}$. We write $\mathcal{T}^{n\times p}$ to denote the set of upper triangular matrices $R \in \mathcal{C}^{n\times p}$. When the matrix coefficients are functions of t, then the notation is modified appropriately, e.g., $\mathbb{C}^{n\times n}(t)$. Capital letters are reserved for matrices, and bold font for vectors, e.g., $\mathbf{y} \in \mathbb{C}^n$. Time derivatives are indicated with a "dot", e.g., $\frac{d\mathbf{y}}{dt} = \dot{\mathbf{y}}$.

1. Introduction.

In a variety of contexts there has recently been widespread interest in decomposition techniques for matrix functions, e.g., see [vLM], [Dav], [Me], [Rh], [GK], [BBMN], [DOR], [Die]. Guided by the linear algebra context, unitary (orthogonal in the real case) factorizations have been utilized in most of these situations. At some level, much of the computational effort is thereby reduced to integrating a skew-Hermitian matrix system whose solution is unitary. The major application we consider is continuous orthonormalization, where a unitary fundamental solution matrix for a differential system is computed (see §2). A number of difficulties integrating these skew-Hermitian initial value problems have been encountered [BDR], [GPL] and special devices have been devised to improve the performance [Me], [GK]. A key dif-

*Received by the editors July 7, 1992; accepted for publication (in revised form) January 15, 1993. This work was supported in part by National Science and Engineering Research Council of Canada grant OGP0008781.

†School of Mathematics, Georgia Institute of Technology, Atlanta, Georgia 30332. The work of this author was supported in part by National Science Foundation grant DMS-9104564.

‡Department of Mathematics and Statistics, Simon Fraser University, Burnaby, British Columbia V5A 1S6, Canada.

§This author was supported in part by National Science and Engineering Research Council of Canada grant OGP0121873.

ficulty is the loss of unitariness in the fundamental solution, the effect of which at times is a halting of the numerical integration for reasons which cannot simply be traced back to the need to meet error tolerances. (This difficulty is observed for both stiff and nonstiff integrators, and explicit and implicit schemes with different stability characteristics.)

Our goal here is to study algorithms for the solution of skew-Hermitian systems which maintain unitariness of the computed solution. There are two ways in which this can be accomplished. One possibility is to use *automatic unitary integrators*: these we define as the standard finite difference integrators which automatically maintain unitariness during integration. We show that standard stability considerations are somewhat inappropriate to describe these so-called unitary integrators. However, it turns out to be simple and straightforward to modify the stability analysis for skew-Hermitian type systems. We show that there are high-order schemes which preserve the unitariness. In particular, the Gauss–Legendre point Runge–Kutta (or simply Gauss RK) methods are the only standard schemes which satisfy this property. An analysis of this in the linear case is given in §3. An extension to the nonlinear case (which includes continuous orthonormalization, where the formulation involves a cubic nonlinearity) is in §4. Central to our analysis is the simple fact that a differentiable unitary matrix $U(t)$ is the solution to a system of the form

$$\dot{U} = H(U,t)\,U,$$

where $H(U,t) \in \mathcal{S}^{n\times n}(t)$ is a skew-Hermitian (possibly nonlinear) matrix operator. The linear case and the continuous orthonormalization case are two instances. In §4 a natural iteration process for the discretization of such systems with Gauss RK schemes allows us to extend the linear results on automatic unitary integration to this nonlinear setting. Convergence results are given for this iteration scheme, which is a particularly efficient implementation of the Gauss RK schemes in our setting.

The second possibility is to use *projected unitary integrators*: these consist of a two-step process in which we first compute an approximation by virtually any scheme, and then project this nonunitary approximation into the manifold of the unitary matrices. We also present these schemes in §4. From a computational perspective, these projected integrators are quite appealing because we can use an explicit scheme for the basic time-stepping procedure, and thus the cost essentially reduces to that of orthonormalizing the computed solution. An algorithmic description of the method and computational considerations are included.

Our general viewpoint is similar to that of much recent work in numerical analysis where emphasis is on preserving the qualitative features of the problem—in this case, on preserving unitariness of the continuous flow at the discrete level. A case close to ours is the integration of Hamiltonian systems with symplectic integrators (for an excellent overview, see [Sa2]).

One of the key benefits in preserving the qualitative features here is that we can show that numerical integration of skew-Hermitian systems with a unitary scheme is equivalent to the exact integration of a perturbed skew-Hermitian system. Quantification of this fact allows one to perform a backward error analysis for the error. Integration with a nonunitary scheme is also equivalent to the exact integration of a perturbed system, but one which is no longer skew-Hermitian (see §§3 and 4).

Practical implementation aspects and numerical comparisons with nonunitary integrators are presented in §5.

It is necessary to stress at this point why and when we think it is important that the computed solution stays unitary. Theoretical and numerical results clearly show

that deterioration of unitariness occurring with nonunitary integrators is of the same order of magnitude as the approximation errors. Therefore, for a small interval of integration this loss is usually not critical. However, our interest originates in problems requiring long time integration, as in the approximation of Lyapunov exponents [GPL], [DRV]. In these cases, maintaining unitariness is critical. Also, the theory for methods using unitary transformations typically relies on 2-norm invariance arguments; to rigorously apply these results to discretizations requires that computed solutions be unitary. Finally, as will be seen, if we do not use a unitary integrator then we are de facto modifying a skew-Hermitian type of problem into a nonskew-Hermitian one, one which is generally either unstable or stiff for large time integration and thus presents all sorts of different difficulties.

2. Motivation and background. Herein we set out to describe the types of problems which have provided the motivation for this work. We begin with a number of simple results.

LEMMA 2.1. *Let* $U \in \mathbb{C}^{n\times n}$. *Then* $U \in \mathcal{U}^{n\times n} \iff U = \exp^S$ *for some* $S \in \mathcal{S}^{n\times n}$.

Proof. ($\Leftarrow$) This is obvious.

($\Rightarrow$) Let $Q \in \mathcal{U}^{n\times n}$ be such that $Q^*UQ = D = \operatorname{diag}(\pm e^{i\lambda_j})$. Therefore, $D = \exp^{i\Lambda}$, with $\Lambda = \operatorname{diag}(\lambda_j, j = 1, \ldots, n)$, and $U = Qe^{i\Lambda}Q^* = e^S, S := iQ\Lambda Q^*$. □

The next lemma is the analogue of this for time varying unitary matrices.

LEMMA 2.2. *Let* $S(t) \in \mathbb{C}^{n\times n}(t)$, *and* $U(t) \in \mathbb{C}^{n\times p}$ *satisfy*

$$\dot{U}(t) = S(t)U(t), \qquad U(0) = U_0, U_0^* U_0 = I. \tag{2.1}$$

Then $U(t) \in \mathcal{U}^{n\times p}(t)$ *for all* $t \iff S(t) \in \mathcal{S}^{n\times n}(t)$.

Proof. Since $U(t) = Z(t)U_0$, where $Z(t)$ is the fundamental solution matrix for the system above (with $Z(0) = I$), we only need to show that $Z(t)$ is unitary $\iff$ $S(t) \in \mathcal{S}^{n\times n}(t)$.

($\Rightarrow$) This is obvious, since $S(t) = \dot{Z}(t)Z^*(t)$, and this right-hand side is trivially skew-Hermitian (just differentiate $Z(t)Z^*(t) = I$).

($\Leftarrow$) Since $\dot{Z}(t) = S(t)Z(t)$, $\dot{Z}^*(t) = -Z^*(t)S(t)$. This may be shown to imply $Z^*(t) = Z^{-1}(t)$, from which the result follows. □

Remark. It follows that any $V \in \mathcal{U}^{n\times n}(t)$ satisfies $\dot{V} = V\widehat{S}$ for some $\widehat{S} \in \mathcal{S}^{n\times n}(t)$, since $U = V^* \in \mathcal{U}^{n\times n}(t)$.

It is of course a well-known fact that modern matrix computation relies heavily on the robustness of unitary factorization techniques. The backbone of these is the QR factorization, whereby a matrix $A \in \mathbb{C}^{n\times p}$ is decomposed into the product

$$A = QR, \tag{2.2}$$

with $Q \in \mathcal{U}^{n\times p}$ and $R \in \mathcal{T}^{p\times p}$. A natural desire is to extend this to compute continuous time varying QR factorizations

$$A(t) = Q(t)R(t) \tag{2.3}$$

of $A(t) \in \mathbb{C}^{n\times p}(t)$. For general square matrices this has been considered by Rheinboldt [Rh]. The related problem of computing a continuous SVD has recently been addressed in [BBMN]. (The issue of obtaining a continuous eigendecomposition for a general $A(t)$ has also received theoretical attention [Ka], [BO].) The key difficulty for all of these is ensuring the smoothness (differentiability) of the factors involved, e.g.,

$Q(t)$. Still, under reasonable conditions these methods are applicable, and the main computational effort then involves the integration of a matrix differential equation whose solution is unitary. The integration of this type of system in such a way that the approximate solution preserves unitariness is the main concern of this paper.

ODE case. Arguably, the outstanding computational difficulty in solving linear differential systems

$$\dot{\mathbf{y}} = A(t)\mathbf{y}(t) + \mathbf{f}(t), \tag{2.4}$$

lies in computing (directly or indirectly) a fundamental solution matrix, i.e., an invertible $Y(t) \in \mathbb{C}^{n\times n}(t)$ satisfying

$$\dot{Y}(t) = A(t)Y(t). \tag{2.5}$$

In this context, it is a well-established fact [AMR] that a desirable approach is to separate $Y(t)$ into its direction and growth components. One possible way to do this is to utilize a *continuous factorization*

$$Y(t) = U(t)R(t), \tag{2.6}$$

where $U^*U = I$, $R(t) \in \mathcal{T}^{n\times n}(t)$. Such a factorization has long been used, both as an analytical tool [Dil] and a computational one [Ab], [Ba]. There is an important distinction to be made between the continuous QR factorization of a fundamental solution matrix $Y(t)$ and that of an arbitrary time dependent matrix. The linear independence of the columns of $Y(t)$ ensures the following.

LEMMA 2.3. *Let $Y(t)$ satisfy* (2.5), *where $A(t)$ is continuous. Then for all t there is a smooth and unique factorization* (2.6) *such that $R(t)$ has positive diagonal elements.*

Proof. The existence and uniqueness of $U(t)$ and $R(t)$ are simple consequences of applying the Gram–Schmidt procedure to the (linearly independent) columns of $Y(t)$. Their differentiability follows similarly by differentiating these Gram–Schmidt equations. □

The differential equations for $U(t)$ and $R(t)$ are easy to construct. Differentiating in (2.6), it follows from (2.5) that $\dot{U}R + U\dot{R} = AUR$, so $\dot{U} = AU - U\dot{R}R^{-1}$. Since $H := U^*\dot{U} \in \mathcal{S}^{n\times n}(t)$ (see Lemma 2.2) and $\dot{R}R^{-1} \in \mathcal{T}^{n\times n}$, we obtain

$$\dot{U} = U\,H(U,t), \tag{2.7}$$

with

$$H(U,t) = U^*AU - \dot{R}R^{-1} \tag{2.8}$$

satisfying

$$(H)_{lm} = \begin{cases} (U^*AU)_{lm}, & \text{if } l > m; \\ i\Im(U^*AU)_{ll}, & \text{if } l = m; \\ -(H)_{ml}, & \text{otherwise.} \end{cases} \tag{2.9}$$

The same set of equations can be derived in the following way. We consider the unitary change of variables $\mathbf{y} = U(t)\mathbf{z}$ for (2.4) such that $\mathbf{z}(t)$ satisfies the modified system

$$\dot{\mathbf{z}} = \widetilde{A}(t)\mathbf{z} + \mathbf{q}(t).$$

The matrix $\widetilde{A}(t) \in \mathcal{T}^{n\times n}$ if and only if $U(t)$ satisfies the so-called *Lyapunov equation*

$$\dot{U} = AU - U\widetilde{A}. \tag{2.10}$$

Moreover, in such case (2.6) holds with $\widetilde{A} = \dot{R}R^{-1}$. This approach of computing the factorization (2.6) is a familiar one in the two-point boundary value problem (BVP) literature [Ab], [Ba], [Dav], [Me], [vLM], where it goes by the name of *continuous orthonormalization.*

Assumptions ensuring unique smooth factors for a continuous SVD of $Y(t)$, i.e., $Y(t) = U(t)\Sigma(t)V^*(t)$ with smooth $U(t)$, $V(t) \in \mathcal{U}^{n\times n}(t)$ and diagonal $\Sigma(t)$, are more restrictive than those of Lemma 2.3. One can guarantee an analytic SVD of $Y(t)$ if $Y(t)$ is analytic (see [BBMN]), but this requires that $A(t)$ is analytic [Har]. To simply have smooth (i.e., differentiable) factors seems to require noncoalescing singular values, a condition which we do not want to impose. In any case, when feasible to compute, an SVD still leads to an equation such as (2.7) for the unitary matrices involved (see [GK] and [Wr]).

In general, system (2.7) can be also formulated for the matrix U^*, rather than U, in which case, $\dot{U}^* = (H(U,t))^*U^*$. For notational convenience, in this work we will focus on the case in which we formally have

$$\dot{U} = H(U,t)\,U, \qquad H(V,t) \in \mathcal{S}^{n\times n}(t) \quad \forall V \in \mathcal{U}^{n\times n}(t) \quad \forall t.$$

Remarks. (i) It is easy to see that a differentiable matrix function $U(t) \in \mathcal{C}^{n\times q}(t)$ satisfies $U(t) \in \mathcal{U}^{n\times q}(t) \iff U(t)$ satisfies a matrix DE

$$\dot{U} = H(U,\dot{U},t)\,U, \qquad U(0) = U_0,\ U_0^*U_0 = I, \tag{2.11}$$

where $H \in \mathcal{S}^{n\times n}(t)$, for all t. This fact can be appreciated by considering, for $q < n$, the extension V of U, $V = (U, W)$, $V \in \mathcal{U}^{n\times n}$. Upon differentiating, $\dot{V} = -V\dot{V}^*V$, so

$$(\dot{U}, \dot{W}) = -(U, W)\begin{pmatrix} \dot{U}^*U & \dot{U}^*W \\ \dot{W}^*U & \dot{W}^*W \end{pmatrix},$$

$$\dot{U} = -(U\dot{U}^* + W\dot{W}^*)U.$$

Now, let $H(U) := -(U\dot{U}^* + W\dot{W}^*)$. Notice that the matrix H in (2.11) is not uniquely defined (unless $q = n$) because the orthonormal basis problem, i.e., the extension of U to V, does not have a unique solution.

(ii) It follows from (i) that if we have $H \in \mathcal{S}^{n\times n}(t)$, but $U_0^*U_0 \neq I$, then $U(t) \notin \mathcal{U}^{n\times q}$; in fact, the defect from unitariness $||I - U^*U||$ does not change with t. Moreover, if we solve a DE of the form (2.11) with $U_0^*U_0 = I$ but with $H \notin \mathcal{S}^{n\times n}(t)$ for all t, then in general $U(t) \notin \mathcal{U}^{n\times q}(t)$.

These observations lead to a better understanding of the stability characteristics of equations (2.1) (with $S(t) \in \mathcal{S}^{n\times n}(t)$) and (2.7) (or (2.10)). They motivate our treatment of numerical schemes in §§3 and 4. Looking ahead, it appears that the usual concepts of stiff or nonstiff equations do not mean much in our context. Of key importance is the need to keep the solution unitary since then the only stepsize restrictions are imposed by accuracy and not stability considerations.

Because of the well-known decoupling and conditioning results for dichotomic linear BVPs (see [Ma], [AMR]), one might suspect that a relationship exists between the dichotomic structure of a linear system and the "stability" of the continuous orthonormalization equations [vLM]. However, as the next example shows, the strength

of the dichotomy generally need not have any impact on the integration (and solution) of (2.7).

Example 2.4. Consider the matrix

$$A(t) = \begin{pmatrix} \alpha\cos(2\beta t) & \beta - \alpha\sin(2\beta t) \\ -\beta - \alpha\sin(2\beta t) & -\alpha\cos(2\beta t) \end{pmatrix} \forall\alpha \quad \text{and} \quad \beta \in \mathbb{R}.$$

The fundamental solution $Y(t)$ such that $Y(0) = I$ can be written as

$$Y(t) = \begin{pmatrix} \cos(\beta t) & \sin(\beta t) \\ -\sin(\beta t) & \cos(\beta t) \end{pmatrix} \begin{pmatrix} e^{\alpha t} & 0 \\ 0 & e^{-\alpha t} \end{pmatrix},$$

from which we have an exponential dichotomy for $\alpha \neq 0$. However, the solution $U(t)$ to (2.7) is always the rotation matrix

$$U(t) = \begin{pmatrix} \cos(\beta t) & \sin(\beta t) \\ -\sin(\beta t) & \cos(\beta t) \end{pmatrix},$$

regardless of α. The matrix H in (2.9) has constant coefficients, viz., $H = \beta\left(\begin{smallmatrix} 0 & 1 \\ -1 & 0 \end{smallmatrix}\right)$. So it is only the speed of rotation β which controls $U(t)$ and not the dichotomy exponent α. Numerically, we expect that only accuracy requirements will limit the stepsize choice when computing $U(t)$.

3. Automatic unitary integrators: linear case. In this section we consider the basic problem of integrating a linear ODE system (2.1) where $S(t) \in \mathcal{S}^{n\times n}(t)$. From Lemma 2.2, we know that the transition matrix $\Phi(t, t_0)$ satisfying $U(t) = \Phi(t, t_0)U(t_0)$, $\Phi(t_0, t_0) = I$, is unitary. A desirable property of any approximation scheme for (2.1) would be that the discrete map advancing the numerical solution maintains unitariness as well. In this section, we investigate those approximation schemes which satisfy this property automatically, i.e., we consider those finite difference discretizations which give unitary solutions. As will be seen shortly, no consistent linear multistep scheme can directly achieve this in general.

DEFINITION 3.1. A consistent one-step scheme of the form

$$U_{k+1} = \Phi_k(h)U_k, \qquad k = 0, 1, 2, \ldots, \tag{3.1}$$

for (2.1) will be called an *automatic unitary integrator*, or *scheme*, if $\Phi_k(h) \in \mathcal{U}^{n\times n}$, for all k and for all h.

We begin our analysis with the model problem

$$\dot{\mathbf{u}} = i\lambda\mathbf{u}, \quad \lambda \in \mathbb{R}, \ \mathbf{u}(0) = \mathbf{u}_0, \quad \mathbf{u}_0^*\mathbf{u}_0 = 1. \tag{3.2}$$

Obviously, a unitary integrator must maintain $\mathbf{u}_k^*\mathbf{u}_k = 1$, for all k, i.e., the solution must remain on the unit sphere. In particular, the numerical scheme cannot introduce artificial solution *growth or dissipation.*

Example 3.2. For the model problem (3.2), consider the following well-known schemes, with $\mu = h\lambda$:

(i) Forward Euler: $\mathbf{u}_{k+1} = (1 + i\mu)\mathbf{u}_k$,

(ii) Backward Euler: $\mathbf{u}_{k+1} = (1/(1 - i\mu))\mathbf{u}_k$,

(iii) Implicit midpoint or trapezoidal rule: $\mathbf{u}_{k+1} = ((2 + i\mu)/(2 - i\mu))\mathbf{u}_k$. Forward and backward Euler produce solutions $\mathbf{u}_k = [r(h)e^{i\tan^{-1}\mu}]^k\mathbf{u}_0$, where $r(h) =$

$(1+\mu^2)^{1/2}$ and $r(h) = (1+\mu^2)^{-1/2}$, respectively, so the solutions spiral outward/inward from the unit sphere with identical phase and the exponential rate $r(h)$. In contrast, the rules in (iii) give solutions $\mathbf{u}_k = e^{i\, k \tan^{-1}(4\mu/(4-\mu^2))}\mathbf{u}_0$ which stay on the unit sphere.

The model problem (3.2) provides useful insight into the potential for the standard integration methods to be automatic unitary integrators. From A-stability type considerations, we realize that the method must have the imaginary axis in its stability domain. In fact, for one-step schemes of the type $\mathbf{u}_1 = R(i\mu)\mathbf{u}_0$ with $\mu = h\lambda$, $\mathbf{u}_1^*\mathbf{u}_1 = 1$ implies that $|R(i\mu)| = 1$; so if $|R(i\mu)| \neq 1$ somewhere along the imaginary axis then the method cannot be unitary.

Next, consider a linear k-step multistep method which for the problem $\dot{\mathbf{u}} = \mathbf{f}(t, \mathbf{u})$ has the form

$$\sum_{j=0}^{k} \alpha_j \mathbf{u}_{m+j} = h \sum_{j=0}^{k} \beta_j \mathbf{f}_{m+j}.$$

For (3.2) one has the characteristic equation

$$\rho(\zeta) - i\mu\sigma(\zeta) = 0, \quad \rho(\zeta) = \sum_{j=0}^{k} \alpha_j \zeta^j, \quad \sigma(\zeta) = \sum_{j=0}^{k} \beta_j \zeta^j.$$

To be unitary the method must have the boundary of its stability region, or "root locus curve" defined by $\theta \to \rho(e^{i\theta})/\sigma(e^{i\theta})$, lie in the right half plane. Moreover, it can be shown that a consistent linear multistep method which is stable on the imaginary axis must be A-stable [Je]. Consequently, if the method is not A-stable then it cannot be unitary. This limits the possibilities to methods of order ≤ 2, and in fact the only one of order 2 which has modulus 1 on the imaginary axis can be seen to be the trapezoidal rule [HW, pp. 259–266]. We will not be concerned with the possible existence of unitary first order linear multistep methods (which are not one-step methods).

Hence, we can reduce our consideration to one-step methods (which include the trapezoidal rule), so that the object of study becomes a rational function $R(z) = P(z)/Q(z)$ advancing the approximate solution of (3.2) from one step to the next by $\mathbf{u}_{k+1} = R(ih\lambda)\mathbf{u}_k$. We have the following lemma.

LEMMA 3.3. *For the model problem* (3.2), *the one-step schemes of the form* $\mathbf{u}_{k+1} = R(i\mu)\mathbf{u}_k$, $\mu = h\lambda$, *are unitary if and only if they are symmetric.*

Proof. After the change h to $-h$, x_k to $x_k + h$, we have that (see [HW]) the method is symmetric if and only if $R(i\mu) = 1/R(-i\mu)$. If $R(i\mu) = 1/R(-i\mu)$, then $|R(i\mu)| = 1$, and hence the method is unitary. Conversely, if the method is unitary, then $|R(i\mu)| = 1$ and therefore $|P(i\mu)| = |Q(i\mu)|$, so that

$$R(i\mu) = \frac{P(i\mu)P(-i\mu)}{Q(i\mu)P(-i\mu)} = \frac{Q(i\mu)Q(-i\mu)}{Q(i\mu)P(-i\mu)} = \frac{1}{R(-i\mu)}. \qquad \square$$

Thus, we now consider the symmetric implicit RK (IRK) schemes, and as is customary, we further restrict attention to schemes which are nonconfluent and have positive weights. That is, if an s-stage IRK is specified by the tableau

$$\begin{array}{c|ccc} c_1 & a_{11} & \dots & a_{1s} \\ \vdots & \vdots & \dots & \vdots \\ c_s & a_{s1} & \dots & a_{ss} \\ \hline & b_1 & \dots & b_s \end{array} \tag{3.3}$$

then $c_i \neq c_j$, $i \neq j$, and $b_i > 0$ for all i. These basically correspond to the collocation schemes [AMR].

Consider now a time dependent model problem

$$\dot{\mathbf{u}}(t) = i\lambda(t)\mathbf{u}(t), \quad \mathbf{u}(0) = \mathbf{u}_0, \quad \mathbf{u}_0^*\mathbf{u}_0 = 1, \tag{3.4}$$

with exact solution satisfying $\mathbf{u}(t)^*\mathbf{u}(t) = 1$ for all t. For a unitary scheme, $\mathbf{u}_k^*\mathbf{u}_k = 1$, for all k and for all h, so the schemes that are not algebraically stable cannot be unitary. In general ([HNW]), this forces the matrix $M = (M_{lm} := b_l a_{lm} + b_m a_{ml} - b_l b_m)_{l,m=1}^s$ to be positive semidefinite. But in fact, the only such symmetric collocation schemes are the Gauss point RK schemes, for which $M = 0$ [AB].

It remains to be shown that these Gauss point RK schemes are indeed unitary. It is convenient to show this indirectly using known results about symplectic integrators for Hamiltonian systems.

A general linear Hamiltonian system has the form $\dot{\mathbf{y}} = M(t)\mathbf{y}$, or

$$\begin{pmatrix} \dot{\mathbf{p}} \\ \dot{\mathbf{q}} \end{pmatrix} = \begin{pmatrix} B & D \\ C & -B^* \end{pmatrix} \begin{pmatrix} \mathbf{p} \\ \mathbf{q} \end{pmatrix}, \tag{3.5}$$

where $D(t)$, $C(t) \in \mathcal{H}^{n\times n}(t)$, $B(t) \in \mathbb{C}^{n\times n}(t)$ and $\mathbf{p}, \mathbf{q} \in \mathbb{C}^n(t)$, and thus $JMJ = M^*$, where $J := \left(\begin{smallmatrix} 0 & -I \\ I & 0 \end{smallmatrix}\right)$. Numerical integrators which preserve the qualitative features of Hamiltonian systems have received considerable recent attention (e.g., see [Sa2]). Specifically, the transition matrix $\Phi(t, t_0)$ for (3.5) is a symplectic transformation, i.e., $\Phi^* J \Phi = J$, and a *symplectic integrator* maintains this property at a discrete level. Thus, for one-step schemes applied to (3.5), if $\mathbf{y}_{k+1} = \Phi_k(h)\mathbf{y}_k$ then $\Phi_k(h)^* J \Phi_k(h) = J$. Since Gauss RK schemes are known to be symplectic [Sa1], we can prove the following.

THEOREM 3.4. *For the general system $\dot{U} = S(t)U$, $U(0) = U_0$, $U_0^* U_0 = I$, with $S(t) \in \mathcal{S}^{n\times n}(t)$, the Gauss RK schemes are unitary.*

Proof. Consider the auxiliary Hamiltonian system

$$\dot{\mathbf{y}} = \begin{pmatrix} \dot{\mathbf{p}} \\ \dot{\mathbf{q}} \end{pmatrix} = \begin{pmatrix} S(t) & 0 \\ 0 & -S^*(t) \end{pmatrix} \begin{pmatrix} \mathbf{p} \\ \mathbf{q} \end{pmatrix} = \begin{pmatrix} S(t) & 0 \\ 0 & S(t) \end{pmatrix} \begin{pmatrix} \mathbf{p} \\ \mathbf{q} \end{pmatrix}.$$

Because of the simple block diagonal structure of the linear system, integration with a Gauss RK scheme gives

$$\mathbf{y}_{k+1} = \begin{pmatrix} \Psi_k(h) & 0 \\ 0 & \Psi_k(h) \end{pmatrix} \mathbf{y}_k.$$

Since the method is symplectic,

$$\begin{pmatrix} \Psi_k^*(h) & 0 \\ 0 & \Psi_k^*(h) \end{pmatrix} \begin{pmatrix} 0 & -I \\ I & 0 \end{pmatrix} \begin{pmatrix} \Psi_k(h) & 0 \\ 0 & \Psi_k(h) \end{pmatrix} = \begin{pmatrix} 0 & -I \\ I & 0 \end{pmatrix},$$

implying that $\Psi_k(h)$ is unitary. Since a Gauss RK scheme for (2.1) has the form $U_{k+1} = \Psi_k U_k$, it is unitary. □

We conclude that for the nonautonomous linear case the Gauss RK schemes produce the only unitary integrators from among the broad class of commonly used difference schemes.

Remarks. (i) The property of being a unitary integrator for a matrix system defined by the mapping Ψ_k is equivalent to $\Psi_k^T \Psi_k = I$ for all k. Although less natural in our context, one could rewrite the matrix system in vector form and interpret the unitary property in terms of preserving bilinear invariants. In fact, in his study of symplectic integrators, Sanz-Serna [Sa1] discusses Gauss RK schemes as they relate to bilinear invariants. From among all known symplectic finite difference schemes, only the Gauss RK schemes are also unitary. Interestingly, we arrive at this result using simple stability arguments whereby classes of methods are successively eliminated. We have been able to do this because all linear skew-Hermitian systems have eigenvalues along the imaginary axis. It could be instructive to take an analogous approach to study symplectic integrators for suitable classes of Hamiltonian systems.

(ii) Studies have been made of stability properties along the imaginary axis in still other contexts, viz., for the integration of hyperbolic systems [JN] and the investigation of spurious solutions for discretizations of IVPs (initial value problems) [IPS].

(iii) A useful property of a unitary integrator is that unitariness is preserved independently of the marching direction. This independence of direction is a property shared by symplectic integrators.

It is now possible to perform an error analysis for unitary schemes. From Lemma 2.1, a discrete map advancing the solution can be written as an exponential of a skew-Hermitian matrix, so a unitary scheme can be interpreted as exactly solving a skew-Hermitian problem. The relationship between this and the original system can be derived as follows.

For the model problem (3.2), a Gauss RK scheme gives

$$\mathbf{u}_{k+1} = R(i\mu)\mathbf{u}_k \qquad (\mu = \lambda h), \tag{3.6}$$

where

$$R(i\mu) = e^{i\phi(\mu)} \tag{3.7}$$

approximates the growth of the exact solution $e^{i\lambda t}$ over one step. Thus, we are solving the modified skew-Hermitian problem

$$\dot{\mathbf{u}} = i(h^{-1}\phi(\mu))\mathbf{u},\ \mathbf{u}_0^*\mathbf{u}_0 = 1. \tag{3.8}$$

For the s-point Gauss RK scheme, $R(i\mu) = P_s(i\mu)/P_s(-i\mu)$, where $P_0 = 1$, $P_1 = 2 + i\mu$, $P_s(i\mu) = 2(2s-1)P_{s-1}(i\mu) + (i\mu)^2 P_{s-2}(i\mu)$, $s = 2, 3, \ldots$. It follows that (with $P_s := P_s(i\mu)$)

$$\phi(\mu) = \tan^{-1}\left(\frac{2\Re(P_s)\,\Im(P_s)}{(\Re(P_s))^2 - (\Im(P_s))^2}\right), \tag{3.9}$$

where $\Re(P_s)$ and $\Im(P_s)$ are the real and imaginary parts of P_s, respectively.

Example 3.5. For the implicit midpoint rule ($s = 1$), equation (3.9) reads $\phi(\mu) = \tan^{-1}(4\mu/(4-\mu^2))$. Taking a Taylor expansion about $\mu = \lambda h = 0$ leads, of course, to the standard error expansion. Explicitly, if $f(\mu) = 4\mu/(4-\mu^2)$ then $f(\mu) = \mu 1/(1 + (i\mu/2)^2) = \mu \sum_{n=0}^{\infty} (\mu/2)^{2n}$. Thus, $\tan^{-1}(f(\mu)) = \mu - \mu^3/12 + \cdots =: \mu + g(\mu)$, and so we have $e^{i\phi(\mu)} = e^{i\mu}e^{ig(\mu)}$, from which the local truncation error is $e^{i\mu}(1 - e^{ig(\mu)}) = +i\mu^3/12 + \mathcal{O}(\mu^5)$.

Example 3.5 hints at the behavior of the global error for general Gauss RK schemes for (3.2). In fact, from (3.7), (3.8), and (3.9) we have that over one step $e^{i\mu}$ is approximated by $e^{i\mu}e^{ig(\mu)}$, where the precise form of $g(\mu)$ depends on the specific integration rule. After k steps, the global error $\mathbf{e}_k := \mathbf{u}(t_k) - \mathbf{u}_k$ satisfies

$$\mathbf{e}_k = e^{ik\mu}(1 - e^{ikg(\mu)})\mathbf{u}_0. \tag{3.10}$$

Thus, the global error is a periodic function, bounded in magnitude by 2. Whenever $kg(\mu)$ is a multiple of 2π, there is *no* error. Of course, in finite precision this might never be exactly true. For example, consider the implicit midpoint rule (see Example 3.5). We expect no error when $k \approx 24\pi m/(h\lambda)^3$, which for realistic values of λ and h only occurs after a very large number of steps.

Another interesting consequence of the error behavior for Gauss RK schemes for (3.2) is the following fact. Rewrite (3.8) as

$$\dot{\mathbf{u}} = i(\psi(\mu)\lambda)\mathbf{u},\ \mathbf{u}_0^*\mathbf{u}_0 = 1, \quad \text{where } \psi(\mu) = \frac{\phi(\mu)}{\mu}. \tag{3.11}$$

For the change of time variable $\tau := t\psi(\mu)$, (3.11) can be rewritten as

$$\frac{d\mathbf{u}}{d\tau} = i\lambda\mathbf{u},\ \mathbf{u}_0^*\mathbf{u}_0 = 1. \tag{3.12}$$

This is precisely the same equation as (3.2) but with a different time scale. Recalling that exact integration of (3.12) is numerical integration of (3.2) with Gauss RK schemes, we see that the numerically computed solution covers the same orbit as the exact solution, but at a different speed. The relevance of these observations is for long time integration, since otherwise only accuracy (hence local error control) is of concern.

These arguments trivially generalize to the constant coefficient matrix case, but the situation is much more complicated for the general variable coefficient case (2.1). One step of the Gauss RK scheme (3.1) gives by Lemma 2.1 $\Phi_k(h) = e^{B_k(h)}$, where $B_k(h) \in \mathcal{S}^{n\times n}$ is unitarily similar to $i\Lambda_k(h)$, with diagonal $\Lambda_k(h) \in \mathbb{R}^{n\times n}$. Although the results for the constant coefficient model problem apply over this interval, so that integration with a Gauss RK scheme is still equivalent to exactly solving a modified skew-Hermitian problem, globally this modified skew-Hermitian problem has a discontinuous right-hand side. For example, for (3.4) the Gauss RK solution

$$\mathbf{u}_{k+1} = R_k(h)\mathbf{u}_k \tag{3.13a}$$

can be seen as the exact solution of

$$\dot{\mathbf{u}} = iG(k,h)\mathbf{u}, \quad G(k,h) = \frac{1}{h}\phi_k(h), \quad \mathbf{u}_0^*\mathbf{u}_0 = 1, \tag{3.13b}$$

with

$$\phi_k(h) = \tan^{-1}(f_k(h)), \quad t_k \le t < t_{k+1}, \quad k = 0, 1, \ldots. \tag{3.13c}$$

While $f_k(h)$ has a similar form to (3.9), it has a k (i.e., time) dependence, e.g., $f_k(h) = 4h\lambda(t_{k+1/2})/(4 - h^2\lambda^2(t_{k+1/2}))$ for the implicit midpoint rule. Thus, the interpretation of global errors is not as transparent as for constant coefficients problems.

One important case in which more can be said is that of periodic variable coefficients. To see this, consider first the periodic skew-Hermitian model problem

$$\dot{\mathbf{u}} = i\lambda(t)\mathbf{u}, \quad \mathbf{u}(0) = \mathbf{u}_0, \quad \mathbf{u}_0^*\mathbf{u}_0 = 1, \quad \lambda(t+T) = \lambda(t), \quad T > 0. \tag{3.14}$$

From Floquet theory [Hal], it is known that

$$\mathbf{u}(t) = p(t)\mathbf{v}(t), \quad p(t) = p(t+T), \quad \dot{\mathbf{v}} = ib\mathbf{v}, \quad b = \frac{1}{T}\int_0^T \lambda(t)dt. \tag{3.15}$$

The following facts are easy to verify.

LEMMA 3.6. (a) *The solution* $\mathbf{u}(t)$ *of* (3.14) *is periodic of period* $T \iff b = \frac{2\pi}{T}m$, $m \in \mathbb{Z}$. (b) *The solution* $\mathbf{u}(t)$ *of* (3.14) *is periodic of period* $\widehat{T}$, *for some* $\widehat{T} \iff b$ *and* $\frac{2\pi}{T}$ *are rationally dependent* (*i.e., there are integers* l *and* m *so that* $lb = \frac{2\pi}{T}m$).

Consider discretization of (3.14) with a Gauss RK scheme and stepsize $h = \frac{T}{M}$, $M \in \mathbb{Z}\,(M \neq 0)$, and the corresponding approximation $\widehat{\mathbf{u}}$ from (3.13a,b,c). From periodicity of $\lambda(t)$ we have that (3.13) in this case is a discontinuous periodic skew-Hermitian problem with period T. Floquet theory gives

$$\widehat{\mathbf{u}}(t) = \widehat{p}(t)\mathbf{w}(t), \quad \widehat{p}(t) = \widehat{p}(t+T), \quad \dot{\mathbf{w}} = i\widehat{b}\mathbf{w}, \quad \widehat{b} = \frac{1}{T}\int_0^T G(k,h)dt. \tag{3.16}$$

We have

$$p(t) = e^{i\int_0^t (\lambda(s)-b)ds}p(0), \quad \widehat{p}(t) = e^{i\int_0^t (G(k,h)-\widehat{b})ds}p(0),$$

so at integer multiples of T the global error is

$$\mathbf{u}(lT) - \widehat{\mathbf{u}}(lT) = p(0)(\mathbf{v}(lT) - \mathbf{w}(lT)), \qquad l \in \mathbb{Z}. \tag{3.17}$$

Thus, at multiples of T the global error only depends on the difference between the solutions of two constant coefficient skew-Hermitian problems. We know how these behave. In particular, $\mathbf{v}(t) = \mathbf{w}(t)$ for t such that $(b-\widehat{b})t$ is a multiple of 2π; since $b-\widehat{b}$ is an $\mathcal{O}(h^{2s-1})$ quantity for an s-stage Gaussian rule, this could still be a long time. The global error can be better characterized in special cases where the assumptions of Lemma 3.6 are satisfied. An important one is when both b and $\widehat{b}$ are 0, for then we need only consider what happens for $0 \leq t < T$. We immediately obtain

$$\mathbf{u}(t) - \widehat{\mathbf{u}}(t) = p(0)\mathbf{v}(0)e^{i\int_0^t \lambda(s)ds}\left(1 - e^{i\int_0^t (G(k,h)-\lambda(s))ds}\right),$$

which displays at worse an $\mathcal{O}(h)$ deterioration in the global error.

The argument for the periodic skew-Hermitian linear system case goes along the following lines. Suppose we have the problem

$$\dot{U}(t) = S(t)U, \quad U(0) = U_0, \quad U_0^*U_0 = I, \quad S^*(t) = -S(t), \quad S(t+T) = S(t), \quad T > 0. \tag{3.18}$$

Since $U \in \mathcal{U}^{n\times n}(t)$, the Floquet theory gives

$$U(t) = P(t)e^{Bt}, \quad , P(t) = P(t+T) \in \mathcal{U}^{n\times n}(t), \quad B = \frac{1}{T}\int_0^T S(t)dt, \tag{3.19a}$$

where $B^* = -B$, and it suffices to assume that

$$U(t) = P(t)e^{i\Lambda t},\ \Lambda = \mathrm{diag}(\lambda_i,\ i = 1,\ldots,n) \in \mathcal{R}^{n\times n}. \tag{3.19b}$$

The results for the model problem (3.14) are now applicable. For example, Lemma 3.6 extends to read: (a) $U(t) = U(t+T) \iff \lambda_i = \frac{2\pi}{T} m_i,\ m_i \in \mathbb{Z}$, and (b) $U(t) = U(t+\widehat{T})$ for some $\widehat{T} \iff$ all λ_i, T are rationally dependent.

The above results are reminiscent of the Hamiltonian case, where one seeks symplectic schemes which give exact solutions to perturbed Hamiltonian systems. The observation that we are exactly solving a perturbed skew-Hermitian linear system may on one hand not appear to be particularly insightful, since it is true by direct construction, simply using the property of unitariness. On the other hand, its implications are nontrivial. It allows us to limit consideration to perturbations belonging to the same class of problems. This is not possible for nonunitary integrators, where the nonskew-Hermitian perturbations lead to numerical instability (cf. Example 3.2 and §5).

These perturbations for a consistent one-step scheme can be quantified. For the model problem (3.2), we still have (3.6), but now

$$R(i\mu) = R_{lp}(i\mu), \qquad R_{lp}(i\mu) := \frac{P_l(i\mu)}{Q_p(i\mu)} = \rho(\mu)e^{i\psi(\mu)}, \tag{3.20}$$

where P_l and Q_p are polynomials of degrees l and p, respectively, and (with P_l and Q_p evaluated at $i\mu$)

$$\begin{aligned} \rho(\mu) &= \left(\frac{(\Re(P_l))^2 + (\Im(P_l))^2}{(\Re(Q_p))^2 + (\Im(Q_p))^2}\right)^{1/2}, \\ \psi(\mu) &= \tan^{-1}\left(\frac{\Im(P)\Re(Q) - \Re(P)\Im(Q)}{\Re(P)\Re(Q) + \Im(P)\Im(Q)}\right). \end{aligned} \tag{3.21}$$

Thus one exactly solves the perturbed problem

$$\dot{\mathbf{u}} = \left[\frac{1}{h}\big(\ln\rho(\mu) + i\psi(\mu)\big)\right]\mathbf{u}, \qquad \mathbf{u}_0^*\mathbf{u}_0 = 1, \tag{3.22}$$

which has the exact solution $\mathbf{u}(t) = \rho(\mu)^{t/h}e^{i\psi(\mu)t/h}\mathbf{u}_0$. The perturbed problem is not skew-Hermitian, and moreover, during integration one never recovers a unitary solution.

Some general error patterns can be observed for nonunitary schemes. For consistent explicit schemes $Q_p(z) = 1$ so that $\rho(\mu) > 1$ (since e^z is approximated to at least $\mathcal{O}(z^2)$ for small z). This means that the unstable problem (3.22) is being integrated, and global errors grow in magnitude in a monotonic fashion. On the other hand, $\rho(\mu) < 1$ for many implicit schemes (e.g., Radau RK schemes), which implies that (3.22) will eventually be stiff and $\mathbf{u}_k$ will approach zero in magnitude. Thus, the behavior observed in Example 3.2 for the Euler rules is somewhat typical. Interestingly, for both implicit and explicit nonunitary schemes, although the computed solution does not resemble the exact solution, the angular component does accurately approximate the exact solution's phase $e^{i\psi(\mu)}$. In other words, if we could separate magnitude and phase for the approximate solution and keep the phase portion, we

could continue the integration, with reasonable hope of maintaining an accurate unitary approximation, as for an automatic unitary integrator. This simple observation provides a motivation for the construction of projected unitary integrators in the next section.

4. Automatic unitary integrators: nonlinear case. Projected unitary integrators.

4.1. Automatic unitary integrators: nonlinear case. Here, we extend the results about automatic unitary integrators from the linear case to a general nonlinear case of the form

(4.1)
$$\dot{U} = H(U,t)\,U(t), \quad U(0) = U_0, \quad U_0^* U_0 = I, \quad H(V,t) \in \S^{n\times n}(t) \quad \forall V \in \mathbb{C}^{n\times q} \quad \forall t.$$

THEOREM 4.1. *All Gauss RK schemes are automatic unitary integrators for* (4.1).

Proof. Consider the s-stage Gauss RK scheme (using (3.3))

$$U_{k+1} = U_k + h\sum_{l=1}^{s} b_l H_{kl} U_{kl}, \tag{4.2a}$$

$$U_{kl} = U_k + h\sum_{j=1}^{s} a_{lj} H_{kj} U_{kj}, \; l = 1,\dots,s, \tag{4.2b}$$

where $H_{kl} := H(U_{kl}, t_k + c_l h)$. For this nonlinear system, consider further the iteration for $m = 0, 1, 2, \dots$:

$$U_{k+1}^{(m+1)} = U_k + h\sum_{l=1}^{s} b_l H_{kl}^{(m)} U_{kl}^{(m+1)}, \tag{4.3a}$$

$$U_{kl}^{(m+1)} = U_k + h\sum_{j=1}^{s} a_{lj} H_{kj}^{(m)} U_{kj}^{(m+1)}, \qquad l = 1,\dots,s, \tag{4.3b}$$

where $H_{kl}^{(m)} := H(U_{kl}^{(m)}, t_k + c_l h)$. By construction, each iteration corresponds to a discretization of some skew-Hermitian linear problem by which Theorem 3.4 yields a unitary solution $U_{k+1}^{(m+1)}$. To prove the theorem it suffices to show that the iteration (4.3b) converges. (Note that the intermediate iterates $U_{k+1}^{(m+1)}$ in (4.3a) are explicitly written simply to emphasize the relationship between the iterates $U_{k1}^{(m+1)}, \dots, U_{ks}^{(m+1)}$, $U_{k+1}^{(m+1)}$ and a linear problem.)

The system (4.2b) can be expressed in the form

$$\big(I_{ns} - h\Omega(W_k)\big)W_k = B_k$$

where I_{ns} is the $ns \times ns$ identity matrix,

$$\Omega = \begin{pmatrix} a_{11}I & \cdots & a_{s1}I \\ \cdots & \cdots & \cdots \\ a_{1s}I & \cdots & a_{ss}I \end{pmatrix} \begin{pmatrix} H(U_{k1}) & & \bigcirc \\ & \ddots & \\ \bigcirc & & H(U_{ks}) \end{pmatrix},$$

$$W_k = \begin{pmatrix} U_{k1} \\ \vdots \\ U_{ks} \end{pmatrix}, \text{ and } B_k = \begin{pmatrix} U_k \\ \vdots \\ U_k \end{pmatrix}.$$

The iteration (4.3b) is then

$$W_k^{(m+1)} = \left(I - h\Omega(W_k^{(m)})\right)^{-1} B_k, \tag{4.4}$$

and convergence follows for h sufficiently small. □

Remark. For continuous orthonormalization, which is the practical situation of concern to us which leads to a nonlinear system of the form (4.1), H is constructed to be skew-Hermitian *using* the unitariness of $U(t)$. Here, the intermediate approximate solution values U_{kl} for the Gauss RK scheme are generally not unitary. To prove unitariness of U_{k+1} we must have each H_{kl} skew-Hermitian, and the approximate H is forced by construction to be skew-Hermitian in (2.9).

In the literature on stiff IVPs, Gauss RK schemes have been regarded as expensive because of the large systems to solve during the Newton iteration. Much effort has gone into finding efficient ways to solve the associated nonlinear systems by alternative means. In our case, for a nonlinearity as in (4.1), the iteration we used in the proof of Theorem 4.1 is also of practical interest, and we have implemented it for our examples in §5. It is easily seen that the contraction constant in the iteration (4.4) is proportional to $h\|H'(U)\|$. Therefore, a small stepsize h is required for convergence whenever $H(U)$ is rapidly varying, while large h is sufficient when $H(U)$ varies slowly (see also §5). Of course, one might consider a different iteration for solving the nonlinear system in (4.2b). The standard choice is Newton's method. Unfortunately, if iteration is not carried to convergence, then the solution need not be a unitary approximation. Moreover, Newton iteration seems generally quite expensive. We can see this for the continuous orthonormalization equations (2.7)–(2.9) as follows: Let

$$(\widehat{H}(U,Z))_{lm} = \begin{cases} (U^*AZ + Z^*AU)_{lm}, & \text{if } l > m; \\ i\Im(U^*AZ + Z^*AU)_{ll}, & \text{if } l = m; \\ -(U^*A^*Z + Z^*A^*U)_{lm}, & \text{otherwise,} \end{cases}$$

and consider the linearization of (2.7) around a solution $U(t)$. Thus, for the first variation $V(t)$ satisfies

$$\dot{V} = U\widehat{H}(U,V) + VH(U),$$

for which there is no obvious exploitable structure computationally.

4.2. Projected unitary integrators. We now consider a class of unitary integrators which can be easily constructed from nonunitary integrators. The basic idea for one-step schemes is as follows: Let U_k be a given unitary approximation to $U(t_k)$ and suppose the approximation $\widehat{U}_{k+1}$ to $U(t_k)$ is computed by a nonunitary scheme. A unitary approximation U_{k+1} can be formed by taking the QR factorization $\widehat{U}_{k+1} = U_{k+1}R_{k+1}$. Proceeding in this fashion we have constructed a unitary integrator. The extension of this approach to multistep schemes is immediate.

The obvious advantage of this procedure is that in theory any nonunitary scheme, even an inexpensive explicit one, suffices. Furthermore, one could intermittently project the computed solution into the set of unitary matrices rather than at every step, since from local error considerations the unitary component of the nonunitary approximation is still approximated accurately for a finite number of steps. More precisely, we have the following.

LEMMA 4.2. *Suppose that* $Z \in \mathbb{C}^{n\times n}$ *and assume that for some* p *and* $h_0 > 0$, $Z^*Z = I + \mathcal{O}(h^p)$ *for* $0 < h < h_0$. *The unique* QR *factorization of* Z *such that* $Z = UR$, $U \in \mathcal{U}^{n\times n}$, $R \in \mathcal{T}^{n\times n}$, *and* $R_{ii} > 0$ *satisfies* $R = I + \mathcal{O}(h^p)$.

Proof. The proof is a simple induction argument for the Gram–Schmidt process. □

To conclude this Section, we look at one possible way to estimate global errors for both automatic and projected unitary integrators. Consider first automatic unitary schemes. Let $U(t,\bar{t})$, $t \geq \bar{t}$, be the exact solution of the DE in (4.1) with initial conditions (ICs) satisfying $U^*(\bar{t})U(\bar{t}) = I$, let U_l be the computed solution at t_l and let $U_l(t)$ be the exact solution of the system satisfying ICs $U_l(t_l) = U_l$. Then, for the computed solution at t_k we have (with $t_0 = 0$)

$$U_k = U(t_k, t_0) + \sum_{j=0}^{k-1} U(t_k, t_{k-j}) E_{k-j} \quad \forall k \text{ and } \forall h, \tag{4.5}$$

where E_{k-j} is the local truncation error matrix. Now consider projected unitary integrators, assuming the projection is done at each step. Let Z_l be the computed unprojected approximations, and let $Z_l = Q_l R_l$ be the QR factorization. For the projected numerical approximations at t_k we have

$$Q_k = U(t_k, t_0) R_1^{-1} \cdots R_k^{-1} + \sum_{j=0}^{k-1} U(t_k, t_{k-j}) F_{k-j} \prod_{l=j}^{0} R_{k-l}^{-1} \quad \forall k \text{ and } \forall h, \tag{4.6}$$

where F_{k-j} is the local truncation error matrix. By Lemma 4.2, each R_l^{-1} is the identity matrix plus terms of the order of the local truncation error, so (4.5) and (4.6) are qualitatively the same. Taking norms in (4.5) to bound the error would show that local errors accumulate at worst linearly. Although no damping can be expected from the (unitary) $U(t_k, t_{k-j})$, the global errors from (4.5) and (4.6) are in fact trivially bounded by 2, since the 2-norm of the difference between *any* two unitary matrices is bounded by 2.

5. Numerical examples. In this section we compare the Gauss unitary integrators of orders two and four, projected unitary integrators, and standard nonunitary integrators on linear and nonlinear problems of skew-Hermitian type. We consider fixed and variable stepsize implementations. As fixed stepsize integrators, we have implemented second- and fourth-order Gauss RK, explicit Runge–Kutta and Adams–Bashforth methods. As variable stepsize integrators, we used the multistep code of Hindmarsh LSODE (with both Adams and Backward Differentiation formulas), and the Runge–Kutta Fehlberg code RKF45. Unless otherwise noted, when using RKF45 and LSODE, both relative and absolute error tolerances are set to 10^{-6}.

Comparison for the integration schemes is done in terms of accuracy (against the exact solution) and preservation of unitariness. Given a matrix U we measure the loss of unitariness as $||I - U^*U||_F$ where $||\cdot||_F$ denotes the Frobenius norm.

Implementation of Gauss methods. The nonlinear iteration (4.3a,b) is implemented to solve (4.2) for the Gauss RK schemes of orders 2 and 4. In our numerical experiments, the convergence test is that $||F_{kl}(U_{k1}, \ldots, U_{ks})||_\infty < 10^{-6}$, where

$$F_{kl}(U_{k1}, \ldots, U_{ks}) \equiv U_{kl} - U_k - h \sum_{j=1}^{s} a_{lj} H_{kj} U_{kj}, \qquad il = 1, \ldots, s.$$

Note that even linear skew-Hermitian problems may require more than one iteration to converge. As initial guess for the nonlinear iteration we simply use the past value

U_k at all Gauss points. A more sophisticated choice would be to use an explicit integrator to provide these initial guesses. This would in fact also yield a local error approximation from which a variable stepsize Gauss scheme could be implemented.

Implementation of projected unitary integrators. To implement a projected unitary integrator it is necessary to project the solution U_{k+1} onto the space of unitary matrices. We do so using the modified Gram–Schmidt process. This is twice as efficient (see [GVL]) as a QR factorization based on Householder transformations; in the latter case, some care must also be exercised to ensure that the R_{ii} are positive. Implementation of projected unitary integrators can be awkward when using existing initial value software. For instance, LSODE uses a Nordsieck array implementation and updating this array with projected values is rather expensive and inconvenient to access. When using RKF45 [SWD], it is necessary to update the code's internally stored function values. While any nonunitary integrator may be used to form a projected unitary integrator, here we restrict attention to explicit nonunitary integrators.

A direct comparison of computational costs between the Gauss RK schemes and the projected integration schemes is difficult to give because of several factors, including stepsize control, frequency of projection, number of iterates of (4.3), etc. However, a rough idea of the relative costs may be obtained by noting that a single step of a fourth-order projected RK scheme has nearly the same cost as a single iteration of a fourth-order Gauss RK scheme.

Example 5.1. Consider the trivial scalar skew-hermitian model problem

$$\begin{cases} \dot{u} = iu \\ u(0) = 1. \end{cases}$$

Figure 1 shows the absolute error using the fourth order Gauss method with a fixed step size and confirms the global error bound in (3.10).

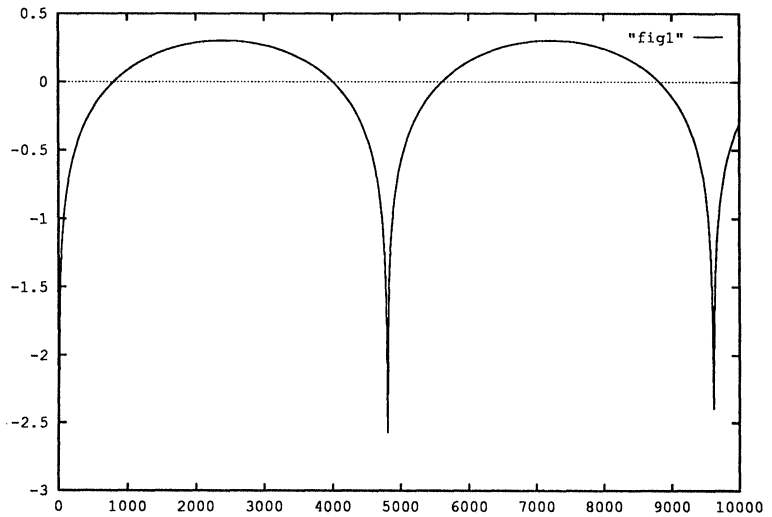

FIG. 1. *Global error* $\log_{10}$ (*Gauss 4,* $h = 1, \dot{u} = iu$)

Example 5.2. Consider the real linear skew-symmetric problem

$$\begin{cases} \dot{U}(t) = \begin{pmatrix} 0 & \beta\sin(\alpha t) \\ -\beta\sin(\alpha t) & 0 \end{pmatrix} U(t), \\ U(0) = I, \end{cases}$$

with exact solution

$$U(t) = \begin{pmatrix} \cos(\phi(t)) & \sin(\phi(t)) \\ -\sin(\phi(t)) & \cos(\phi(t)) \end{pmatrix},$$

where $\phi(t) = \frac{\beta}{\alpha}(1 - \cos(\alpha t))$.

In Table 1 the accuracy and unitariness of the computed solution are compared for the following fixed step methods: the second-order Gauss scheme (Gauss 2), the fourth-order Gauss scheme (Gauss 4), the second-order (Heun's method) and the fourth-order (the classical) Runge–Kutta schemes (RK2 and RK4), the second- and fourth-order Adams–Bashforth methods (AB2 and AB4), and the corresponding projected unitary integrators (PRK2, PRK4, PAB2, and PAB4). In Table 2 various variable stepsize codes are compared, viz., an unprojected and a projected Runge–Kutta Fehlberg code (RKF45 and PRKF45, respectively) and LSODE with maximum order of 2 and 4 (LSODE 2 and LSODE 4) using both Adams and BDF formulas. All computations are for $0 \leq t \leq T \equiv 1000$. Figure 2 gives a log plot of the global error for Gauss 2 with $\alpha = \beta = 1$.

TABLE 1

Example 2: Fixed Step Methods ($h = 0.1, T = 1000, \alpha = \beta = 1$)		
Method	Global Error	Unitary Error
Gauss 2	3.3E-04	2.2E-14
Gauss 4	6.0E-07	1.0E-14
RK 2	8.1E-02	2.4E-01
RK 4	4.3E-05	1.2E-04
AB 2	2.5E-01	8.1E-01
AB 4	1.4E-02	4.0E-02
PRK 2	2.8E-03	4.4E-16
PRK 4	7.9E-07	4.4E-16
PAB 2	1.3E-02	6.3E-16
PAB 4	3.1E-04	6.3E-16

TABLE 2

Example 2: Variable Step Methods ($T = 1000, \alpha = \beta = 1$)		
Method	Global Error	Unitary Error
RKF45	1.5E-03	4.2E-03
PRKF45	4.6E-05	4.4E-16
LSODE 2 (Adams)	9.5E-03	2.6E-02
LSODE 2 (BDF)	7.8E-03	2.2E-02
LSODE 4 (Adams)	5.9E-03	1.7E-02
LSODE 4 (BDF)	2.2E-03	6.2E-03

Example 5.3. Consider now the linear skew-Hermitian system

$$\begin{cases} \dot{U}(t) = \begin{pmatrix} i\cos(\alpha t) & \beta t \\ -\beta t & i\cos(\alpha t) \end{pmatrix} U(t) \\ U(0) = I, \end{cases}$$

with exact solution

$$U(t) = \exp(i\theta(t)) \begin{pmatrix} \cos(\phi(t)) & \sin(\phi(t)) \\ -\sin(\phi(t)) & \cos(\phi(t)) \end{pmatrix},$$

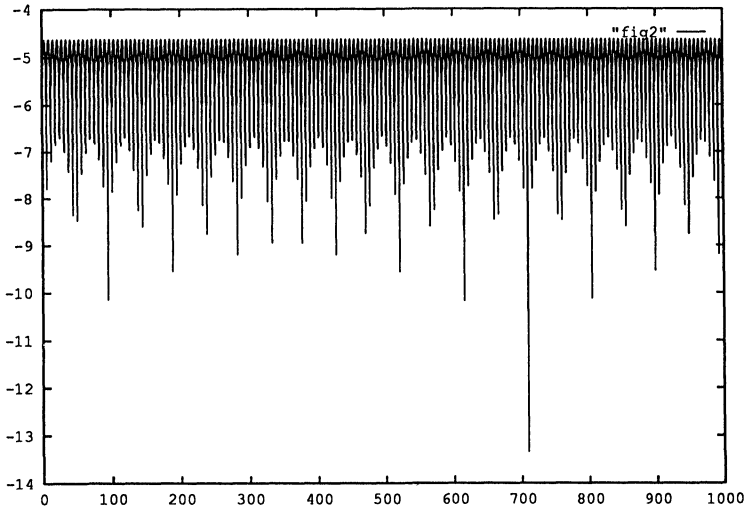

FIG. 2. *Global error* $\log_{10}$ (*Gauss 4*, $h = 0.25, \alpha = \beta = 1$).

where $\theta(t) = \frac{1}{\alpha}\sin(\alpha t)$ and $\phi(t) = \frac{\beta}{2}t^2$.

In Tables 3 and 4 we summarize the global error and loss of unitariness in the computed solution for various integrators with fixed stepsize and for $0 \leq t \leq T \equiv 1000$.

TABLE 3

Example 3: Fixed Step Methods ($h = 0.1, T = 1000, \alpha = 1, \beta = 0$)		
Method	Global Error	Unitary Error
Gauss 2	1.9E-04	2.5E-14
Gauss 4	3.3E-07	1.1E-14
RK 2	8.1E-02	2.3E-01
RK 4	4.3E-05	1.2E-04
PRK 2	1.6E-03	6.3E-16
PRK 4	4.1E-07	6.3E-16

TABLE 4

Example 3: Fixed Step Methods ($h = 0.1, T = 1000, \alpha = 1, \beta = 1$)		
Method	Global Error	Unitary Error
Gauss 2	1.99	6.0E-14
Gauss 4	1.99	5.1E-12
RK 2	Failed	Failed
RK 4	Failed	Failed
PRK 2	1.99	6.4E-16
PRK 4	1.99	6.5E-16

Figures 3 and 4 are log plots of the global error for different values of α and β. Figure 3 shows that the global error remains small for $\alpha = 1$ and $\beta = 0$; this is expected, since the coefficient matrix is periodic with mean zero (cf. §3). Figure 4 shows that the global error is rather erratic and realizes its maximum possible value of 2 for $\alpha = 1$ and $\beta = 1$.

Example 5.4. Consider the linear system

$$\epsilon\ddot{y} - \gamma\dot{y} + ty = 0, \qquad 0 < \epsilon << 1,$$

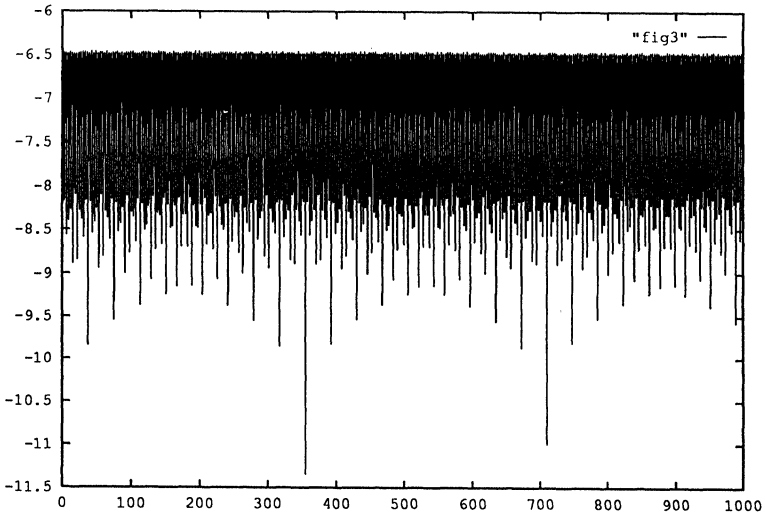

FIG. 3. *Global error* log_{10} (*Gauss 4*, $h = 0.1$, $\alpha = 1, \beta = 0$).

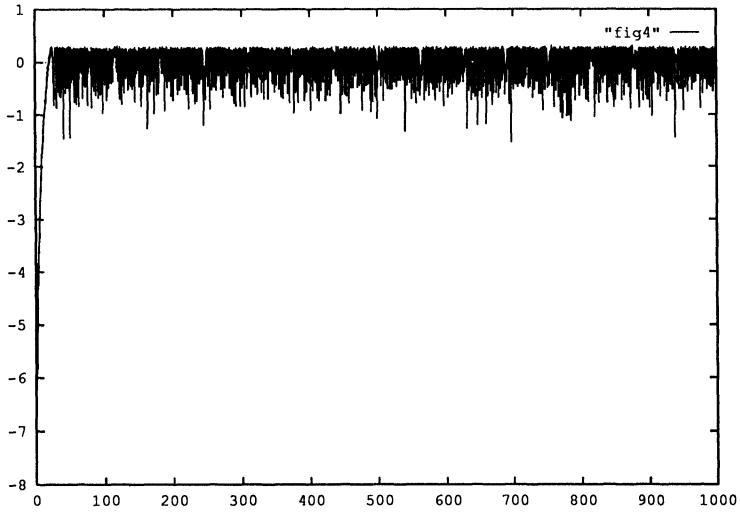

Fig. 4. *Global error* log_{10} (*Gauss 4*, $h = 0.1$, $\alpha = 1, \beta = 1$).

for $-1 \leq t \leq +1$ and $\gamma > 0$ and the continuous orthonormalization equations (2.7)–(2.9). This example corresponds to a two-point boundary value problem with a solution space which changes from oscillatory for $t < 0$ to dichotomic for $t > 0$. The interest is to see whether there is any impact on the integration for the unitary matrix $U(t)$. None is observed (see Table 5).

TABLE 5

Example 4 ($\epsilon = 0.01, \gamma = 1, h = 0.01$)	
Method	Unitary Error
Gauss 2	6.6E-16
Gauss 4	3.9E-15
RK 2	2.7E-01
RK 4	1.4E-02
PRK 2	4.7E-16
PRK 4	6.3E-16
RKF45	2.9E-06
PRKF45	3.1E-16

Example 5.5. Consider the linear matrix differential equation

$$\dot{Y}(t) = \begin{pmatrix} -10 & 0 \\ 20 & 10 \end{pmatrix} Y(t).$$

When using continuous orthonormalization to find $U(t)$ with $U(0) = \left(\begin{smallmatrix} 1 & 0 \\ 0 & 1 \end{smallmatrix}\right)$ the solution quickly approaches $V = \left(\begin{smallmatrix} 0 & 1 \\ 1 & 0 \end{smallmatrix}\right)$ (see [vLM]). From the summary in Table 6 we see that although RKF45 initially takes small steps in order to maintain accuracy, eventually it takes reasonably large time steps. In fact, the maximum step-size is basically the same value for several different error tolerances. In this respect, the time integration for $U(t)$ is not a stiff IVP.

TABLE 6

Example 5 (PRKF45, $T = 10$)		
Local Error Tolerance	Minimum Step Size	Maximum Step Size
1.E-04	1.1E-02	3.2E-01
1.E-06	6.2E-03	2.3E-01
1.E-08	2.4E-03	2.4E-01

Acknowledgments. We are grateful to Timo Eirola and Andrew Stuart for helpful remarks on an earlier version of this paper.

REFERENCES

[Ab] A. A. ABRAMOV, *On the transfer of boundary conditions for systems of ordinary linear differential equations (a variant of the dispersive method)*, USSR Comp. Math. & Math. Phys., 1 (1962), pp. 617–622.

[AB] U. ASCHER AND G. BADER, *Stability of collocation at Gaussian points*, SIAM J. Numer. Anal., 23 (1986), pp. 412–422.

[AMR] U. ASCHER, R. M. MATTHEIJ, AND R. D. RUSSELL, *Numerical Solution of Boundary Value Problems for ODEs*, Prentice-Hall, Englewood Cliffs, NJ, 1988.

[Ba] N. S. BAKHVALOV, *Numerical methods*, MIR, Moscow, 1977; Russian ed., 1975.

[BBMN] A. BUNSE-GERSTNER, R. BYERS, V. MEHRMANN, AND N. K. NICHOLS, *Numerical computation of an analytic singular value decomposition by a matrix valued function*, Numer. Math., 60 (1991), pp. 1–40.

[BO] J. V. BURKE AND M. J. OVERTON, *Stable perturbations of nonsymmetric matrices*, Lin. Alg. and Appl., 171 (1992), pp. 249–274.

[BDR] S. BRAMLEY, L. DIECI, AND R. D. RUSSELL, *Numerical solution of eigenvalue problems for linear boundary value ODEs*, J. Comput. Phys., 94 (1991), pp. 382–402.

[Dav] A. DAVEY, *An automatic orthonormalization method for solving stiff BVPs*, J. Comput. Phys., 51 (1983), pp. 343–356.

[Die] L. DIECI, *Numerical integration of the differential Riccati equation and some related issues*, SIAM J. Numer. Anal., 29 (1992), pp. 781–815.

[DOR] L. DIECI, M. R. OSBORNE, AND R. D. RUSSELL, *A Riccati transformation method for solving linear BVP* I: *theoretical aspects*; II: *computational aspects*, SIAM J. Numer. Anal., 25 (1988), pp. 1055–1092.

[DRV] L. DIECI, R. D. RUSSELL, AND E. S. VAN VLECK, *On the computation of Lyapunov exponents*, submitted.

[Dil] S. P. DILIBERTO, *On systems of ordinary differential equations*, in Contributions to the Theory of Nonlinear Oscillations Ann. of Math. Stud. 20, Princeton Univ. Press, Princeton, NJ, 1950, pp. 1–38.

[GPL] K. GEIST, U. PARLITZ, AND W. LAUTERBORN, *Comparison of different methods for computing Lyapunov exponents*, Prog. Theoret. Phys., 83 (1990), pp. 875–893.

[GVL] G. H. GOLUB AND C. F. VAN LOAN, *Matrix Computations*, Johns Hopkins Univ. Press, Baltimore, MD, 1983.

[GK] J. M. GREENE AND J-S. KIM, *The calculation of Lyapunov spectra*, Phys. D, 24 (1987), pp. 213–225.

[HNW] E. HAIRER, S. P. NORSETT, AND G. WANNER, *Solving Ordinary Differential Equations* I: *Nonstiff Problems*, Springer-Verlag, Berlin, 1987.

[HW] E. HAIRER AND G. WANNER, *Solving Ordinary Differential Equations* II: *Stiff Problems and Differential-Algebraic Equations*, Springer-Verlag, Berlin, 1991.

[Hal] J. K. HALE, *Ordinary Differential Equations*, Krieger Publishing Co., Malabar, Florida, 1980.

[Har] P. HARTMAN, *Ordinary Differential Equations*, John Wiley & Sons, New York, 1964.

[IPS] A. ISERLES, A. T. PEPLOW, AND A. M. STUART, *A unified approach to spurious solutions introduced by time discretization. Part* I: *basic theory*, SIAM J. Numer. Anal., 28 (1991), pp. 1723–1751.

[Je] R. JELTSCH, *Stability on the imaginary axis and A-stability of linear multistep methods*, BIT, 18 (1978), pp. 170–174.

[JN] R. JELTSCH AND O. NEVANLINNA, *Stability of semidiscretizations of hyperbolic problems*, SIAM J. Numer. Anal., 20 (1983), pp. 1210–1218.

[Ka] T. KATO, *Perturbation Theory for Linear Operators*, Springer-Verlag, Berlin, 2nd ed., 1976.

[vLM] P. M. VAN LOON AND R. M. M. MATTHEIJ, *Stable continuous orthonormalization techniques for linear boundary value problems*, J. Austr. Math. Soc. Ser. B, 29 (1988), pp. 282–292.

[Ma] R. M. M. MATTHEIJ, *Decoupling and stability of algorithms for boundary value problems*, SIAM Rev., 27 (1985), pp. 1–44.

[Me] G. H. MEYER, *Continuous orthonormalization for boundary value problems*, J. Comput. Phys., 62 (1986), pp. 248–262.

[Rh] W. RHEINBOLDT, *On the computation of multi-dimensional solution manifolds of parametrized equations*, Numer. Math., 53 (1988), pp. 165–181.

[Sa1] J. M. SANZ-SERNA, *Runge–Kutta schemes for hamiltonian systems*, BIT, 28 (1988), pp. 877–883.

[Sa2] ———, *Symplectic integrators for Hamiltonian problems: An overview*, Acta Numerica, 1 (1991), pp. 243–286.

[SWD] L. F. SHAMPINE H. A. WATTS, AND S. M. DAVENPORT, *Solving non-stiff ODEs—the state of the art*, SIAM Rev., 18 (1976), pp. 376–411.

[Wr] K. WRIGHT, *Differential equations for the analytic singular value decomposition of a matrix*, Numer. Math., 63 (1992), pp. 283–296.