

# ON THE ERROR IN QR INTEGRATION \*

LUCA DIECI<sup>†</sup> AND ERIK S. VAN VLECK<sup>‡</sup>

**Abstract.** An important change of variables for a linear time varying system  $\dot{x} = A(t)x$ ,  $t \geq 0$ , is that induced by the QR-factorization of the underlying fundamental matrix solution:  $X = QR$ , with  $Q$  orthogonal and  $R$  upper triangular (with positive diagonal). To find this change of variable, one needs to solve a nonlinear matrix differential equation for  $Q$ . Practically, this means to find a numerical approximation to  $Q$  by using some appropriate discretization scheme, whereby one attempts to control the local error during the integration. Our contribution in this work is to obtain global error bounds for the numerically computed  $Q$ . These bounds depend on the local error tolerance used to integrate for  $Q$ , and on structural properties of the problem itself, but not on the length of the interval over which we integrate. This is particularly important, since –in principle–  $Q$  may need to be found on the half-line  $t \geq 0$ .

**Key words.** QR methods, Orthogonal Integration, Lyapunov exponents, integral separation.

**AMS subject classifications.** 65L

**Notation.** An  $(n \times n)$  real matrix  $X$  is indicated with  $X \in \mathbb{R}^{n \times n}$ .  $\text{diag}(X)$  is the matrix made up by the diagonal part of  $X$ , the rest being all 0's,  $\text{upp}(X)$  is the matrix made up by the upper triangular part of  $X$ , the rest being all 0's, and  $\text{low}(X)$  is the matrix made up by the strictly lower triangular part of  $X$ , the rest being all 0's. The default norm we consider is the 2-norm of vectors and the induced norm for matrices.

**1. Introduction.** Consider the homogeneous nonautonomous linear differential equation

$$(1.1) \quad \dot{x}(t) = A(t)x(t), \quad t \geq 0,$$

where  $A$  is a bounded function taking values in  $\mathbb{R}^{n \times n}$ . Equation (1.1) appears pervasively in the study of dynamical systems. E.g., it is the equation we end up with when we study variation with respect to the initial conditions, or parameters, of a nonlinear system. Therefore, it is the problem we have to face when we do general stability analyses for trajectories of a dynamical system; e.g., for periodic or for chaotic trajectories. Moreover, (1.1) is also the problem at hand during a Newton process to solve general nonlinear differential systems, a process often advocated for solving boundary value problems. Alas, in spite of its apparent simplicity, numerical investigation of (1.1) is extremely hard, since the solution structure depends on the fundamental matrix solution. Unquestionably, the problem is certainly conceptually and computationally simpler if  $A$  happened to be triangular. For this reason, techniques which find an orthogonal change of variable to triangular structure have been studied by several researchers for a long time; e.g., see [6, 8, 18]. Our own interest in these techniques originates with methods to approximate Lyapunov exponents of dynamical systems, a feat which is greatly simplified when the system is brought in upper triangular structure; e.g., see [9, 11, 12].

Of course, the factor  $Q$  which performs the change of variables has to be found numerically, and this itself is not easy since  $Q$  satisfies a nonlinear matrix differential equation. Thus, to find  $Q$ , one must approximate the solution of this nonlinear matrix equation in some appropriate way. In practice, this means that we will control the local errors while approximating  $Q$ , a fact which generally does not guarantee that  $Q$  will be approximated accurately, that is that the global error in our approximation will stay small. Our contribution in this work is to provide accurate bounds on the **global** error when finding  $Q$ : Our bounds will depend on the **local** error tolerance and on the coefficient function  $A$ , but not on the length of the interval over which we approximate  $Q$ . Our result is somewhat atypical and is important. It is atypical, because, even though  $Q$  lies in a compact space, usually one does not obtain accurate global error bounds (on arbitrarily long intervals) except for contractive problems, and our problem is not contractive. It is important, because –used in conjunction with standard techniques to approximate Lyapunov exponents– it can be used to obtain global error bounds on the computed Lyapunov exponents of linear time varying system, as well as global errors on other spectral quantities.

---

\*This work was supported in part under NSF Grants DMS-0139895, DMS-0139824, and DMS-0513438.

<sup>†</sup>School of Mathematics, Georgia Institute of Technology, Atlanta, Georgia 30332 ([dieci@math.gatech.edu](mailto:dieci@math.gatech.edu)).

<sup>‡</sup>Department of Mathematics, University of Kansas, Lawrence, Kansas 66045 ([evanvleck@math.ku.edu](mailto:evanvleck@math.ku.edu)).

The way we will obtain global error bounds for the computed  $Q$  is in itself interesting, and apparently new. Our main idea is to combine two types of error analyses: A backward error analysis guaranteeing that the computed  $Q$  factor gives a transformation to nearly triangular form, and a forward error analysis guaranteeing that for this nearly triangular problem there is a near the identity orthogonal transformation reducing it further to triangular structure. Combining these two ingredients, we will obtain the sought result. Oversimplifying it, let us sketch the basic idea which has guided us:

- We want to express  $X = QR$ ,  $Q$  orthogonal,  $R$  upper triangular with positive diagonal.
  - If we had  $Q$ , then  $R$  would satisfy a triangular system  $\dot{R} = \tilde{B}R$ .
  - Suppose that instead of  $Q$  we compute (*backward error result*) an orthogonal  $Q_c$ , which gives  $X = Q_c \hat{R}$ , with  $\hat{R} = (B + F)\hat{R}$ , with  $B$  triangular and  $F$  of small norm ( $F$  not triangular).
  - Suppose also that we write  $\hat{R} = VU$ , with  $V$  orthogonal and  $U$  upper triangular with positive diagonal. Then, we have  $X = Q_c VU$ , and so, by uniqueness,  $R = U$  and  $Q_c V = Q$ .
- If we now show that  $V \approx I$  (*forward error result*), then we will infer that  $Q_c \approx Q$  (*global error result*).

A plan of the paper is as follows. In the remainder of this Introduction we review the basic change of variables  $X = QR$  and the differential equations satisfied by  $Q$  and  $R$ . In Section 2 we recall the key backward error statement which we proved in [13]. In Section 3 we give in a concise way the global error statement result, and in Section 4 we give details of a systematic way to obtain sharp bounds on the quantities appearing in the error bound. In Section 5 we illustrate our results on an Example. Conclusions are in Section 6, which include a remark on the modifications needed to handle the case in which we only have a “reduced” QR factorization, that is when  $X$  only comprises a subset of columns of the fundamental matrix solution.

We now consider the differential equations governing the evolution of the  $Q$  and  $R$  factors in the QR factorization of  $X$ . Presently,  $X$  is a fundamental matrix solution for (1.1):  $\dot{X} = A(t)X$ ,  $X(0) = X_0$  invertible. Let  $X_0 = Q_0 R_0$  be the unique QR-factorization of  $X_0$  with the diagonal of  $R_0$  being positive.

Differentiating the relation  $X = QR$  one obtains  $\dot{Q}R + Q\dot{R} = A(t)QR$  and multiplying by  $Q^T$  on the left we obtain the equation for  $R$ :

$$(1.2) \quad \dot{R} = \tilde{B}(t)R, \quad R(0) = R_0,$$

where we have set

$$(1.3) \quad \tilde{B}(t) := Q^T(t)A(t)Q(t) - Q^T(t)\dot{Q}(t).$$

Let us formally set  $S := Q^T\dot{Q}$ . Since  $R$  has to be upper triangular, we must have  $\tilde{B}$  upper triangular, which leads to

$$(1.4) \quad \dot{Q} = QS(Q, A(t)), \quad t \geq 0,$$

where

$$(1.5) \quad S(Q(t), A(t))_{ij} = \begin{cases} (Q^T(t)A(t)Q(t))_{ij}, & i > j, \\ 0, & i = j, \\ -(Q^T(t)A(t)Q(t))_{ji}, & i < j. \end{cases}$$

In particular, we notice that if  $Q$  is known, then  $R$  satisfies (1.2), and we also notice that  $S$  is linear in  $A$ . Furthermore, in light of (1.5), for the entries of  $\tilde{B}$  we have  $\tilde{B}_{ij} = (Q^T A Q)_{ij} + (Q^T A Q)_{ji}$  for  $i < j$  and  $\tilde{B}_{ii} = (Q^T A Q)_{ii}$ , that is  $\tilde{B} = \text{upp}(Q^T A Q) + (\text{low}(Q^T A Q))^T$ .

The above derivation of the equations for the QR-factorization of  $X$  has been obtained many times before, and specific attention has been paid in recent years to techniques which maintain orthogonality while approximating the factor  $Q$ . A sample of relevant references include [2, 3, 7, 8, 10, 16]. We are not going to review these works in details, because the precise way in which the approximation for  $Q$  is obtained is not relevant to our main scope here, which is to derive global error bounds for the obtained approximations to  $Q$ . What is relevant is that the obtained approximations be orthogonal at the grid-points found during numerical integration of (1.4), a fact which the schemes proposed in the above cited works do achieve.

**2. Background.** Suppose we are seeking the factorization  $X(t_k) = Q(t_k)R(t_k)$ ,  $k = 0, 1, 2, \dots$ . In other words, we are looking for the change of variables, the factor  $Q$  in the QR-factorization of  $X$ , at the grid-points  $0 = t_0 < t_1 < \dots$ . Practically, the gridpoints  $\{t_k\}$  may have been found during numerical integration of (1.4) by any of the schemes in the previously cited works. Alternatively, we can always think of indirectly having found approximations to  $Q$  by directly seeking the QR factorization of  $X(t_k)$ , as follows. Write

$$(2.1) \quad X(t_k) = \Phi(t_k, t_{k-1}) \dots \Phi(t_2, t_1) \Phi(t_1, 0) X_0,$$

where

$$(2.2) \quad \dot{\Phi}(t, t_{j-1}) = A(t)\Phi(t, t_{j-1}), \quad \Phi(t_{j-1}, t_{j-1}) = I, \quad t_{j-1} \leq t \leq t_j, \quad j = 1, 2, \dots, k.$$

Then, for  $j = 1, 2, \dots, k$ , recursively consider (*discrete QR technique*)

$$(2.3) \quad \begin{aligned} \dot{\Psi}(t, t_{j-1}) &= A(t)\Psi(t, t_{j-1}), \quad \Psi(t_{j-1}, t_{j-1}) = Q(t_{j-1}) \\ \text{and factor } \Psi(t_j, t_{j-1}) &= Q(t_j)R(t_j, t_{j-1}), \end{aligned}$$

where  $Q(t_j)$  are orthogonal and  $R(t_j, t_{j-1})$  are upper triangular with positive diagonal. So, we have the QR factorization of  $X(t_k)$

$$(2.4) \quad X(t_k) = Q(t_k)R(t_k, t_{k-1}) \dots R(t_2, t_1)R(t_1, t_0)R(t_0).$$

If we adopt this point of view, the error we commit in finding  $Q$  is inherited from the error we do when approximating the transition matrices  $\Phi(t_j, t_{j-1})$ ,  $j = 1, 2, \dots, k$ .

Notice that taking this point of view, we have expressed  $R(t_k)$  as product of local triangular transition matrices:

$$R(t_k) = R(t_k, t_{k-1}) \dots R(t_2, t_1)R(t_1, t_0)$$

where each of these triangular transition matrices is the same as the solution of

$$\dot{R}(t, t_{j-1}) = \tilde{B}(t)R(t, t_{j-1}), \quad R(t_{j-1}, t_{j-1}) = I, \quad t_{j-1} \leq t \leq t_j, \quad j = 1, 2, \dots, k,$$

where  $\tilde{B}$  is given in (1.3).

Now, we cannot hope to be able to obtain the exact factors  $Q(t_k)$  (and  $R(t_k)$ ). Still, let us assume that the obtained numerical approximations to the  $Q(t_k)$ 's, call then  $Q_k$ 's, are orthogonal. The key fact, which we proved in [13], is the following: "By using either direct integration of (1.4) or having indirectly approximated  $Q$  via the discrete QR technique, with a numerical realization of the change of variables  $X = QR$ , we are obtaining a numerical approximation to  $X(t_k)$ , call it  $X_k$ , and to the triangular transition matrices  $R(t_k, t_{k-1})$ , call these  $R_k$ , so that we have

$$(2.5) \quad X_k = Q_k R_k R_{k-1} \dots R_2 R_1 R(t_0), \quad k = 1, 2, \dots,$$

and at the same time

$$(2.6) \quad X_k = Q(t_k)[R(t_k, t_{k-1}) + E_k] \dots [R(t_2, t_1) + E_2][R(t_1, t_0) + E_1]R(t_0), \quad k = 1, 2, \dots,$$

where  $Q(t_k)$  is the exact  $Q$ -factor at  $t_k$  and the triangular transitions  $R(t_j, t_{j-1})$  are also the exact ones. Moreover, the factors  $E_j$ ,  $j = 1, \dots, k$ , are bounded in norm by the local error committed during integration of the relevant differential equations; see [13, Theorems 3.1 and 3.16]."

We will henceforth simply write

$$(2.7) \quad \|E_j\| \leq \eta, \quad j = 1, 2, \dots,$$

and stress that  $\eta$  is computable, in fact controllable, in terms of local error tolerances.

Furthermore, close inspection of the error terms  $E_j$ ,  $j = 1, \dots, k$ ,  $k = 1, 2, \dots$ , allowed us to obtain a backward error result, which we summarize below. For details, we refer to the original work (see [13, Theorem 3.12]), here we are content with a useful rephrasing of this result.

**SUMMARY 2.1.** *With a numerical realization of the QR change of variables, either having directly integrated (1.4) or indirectly through the numerical realization of the discrete QR technique, we do not obtain the exact transformation to the triangular form (1.2-1.3), but rather find an orthogonal change of variable to the perturbed triangular system*

$$(2.8) \quad \dot{\hat{R}} = (B(t) + F(t))\hat{R}, \quad t \geq 0,$$

where  $B$  is the piecewise constant (and triangular) function given by

$$(2.9) \quad B(t) := \frac{1}{t_{j+1} - t_j} \log(R(t_{j+1}, t_j)), \quad t_j \leq t < t_{j+1},$$

and  $F$  is bounded as

$$(2.10) \quad \sup_{t \geq 0} \|F\| \leq c\eta + O(\eta^2) =: \delta.$$

For the sake of completeness, we remark that in (2.10), the bounds on the norm of  $F$  are obtained locally, on each subinterval  $[t_{j-1}, t_j]$ ,  $j = 1, 2, \dots$ , so that one really has  $\sup_{t_{j-1} \leq t \leq t_j} \|F\| \leq c_j\eta + O(\eta^2)$ , and the main contribution to the magnification factor  $c_j$  is given by the departure from normality of the exact triangular transition factors  $R(t_j, t_{j-1})$ . Indeed, at first order in TOL, we have

$$(2.11) \quad \sup_{t_{j-1} \leq t \leq t_j} \|F\| \approx \text{TOL}(1 + \kappa_{j-1}h_{j-1}), \quad h_{j-1} = t_j - t_{j-1},$$

where TOL is the local error for the obtained approximation to the transition matrix, and  $\kappa_{j-1}$  is the departure from normality of  $R(t_j, t_{j-1})$ . In any case, we stress once more that the bounds on the norm of  $F$  are computable.

*Remark 2.1.* In order to obtain (2.10), in [13], we needed to have a certain condition satisfied; see [13, Assumption 3.5]. This amounted to the requirement that

$$\text{TOL}[\min(1, \min_{1 \leq i \leq n} \exp(\int_{t_{j-1}}^{t_j} B_{ii}(s)ds))]^{-1} < 1.$$

In practice, this means that one may need to have the stepsizes  $h_j := t_j - t_{j-1}$ ,  $j = 1, 2, \dots$ , sufficiently small.

**3. Global Error Bounds for Q.** Next, consider the unperturbed and perturbed triangular systems

$$(3.1) \quad \dot{R} = B(t)R, \quad R(0) = R_0 \quad \text{and} \quad \dot{\hat{R}} = [B(t) + F(t)]\hat{R}, \quad \hat{R}(0) = R_0,$$

where we can assume that  $\sup_{t \geq 0} \|F(t)\| \leq \delta$ .

In [14], we proved (Lemma 3.1 below) that there is an orthogonal change of variables, close to the identity, taking the perturbed triangular system to triangular form. The proof we gave used global (and fairly crude) norm estimates and proceeded as follows.

First, write  $R = R_D + R_U$ , where  $R_D = \text{diag}(R)$  and  $R_U = \text{upp}(R)$ , so that  $R = (I + R_U R_D^{-1})R_D =: ZR_D$ . Accordingly, we have the unperturbed and perturbed diagonal systems

$$(3.2) \quad \dot{R}_D = D(t)R_D, \quad \dot{\hat{R}}_D = [D(t) + E(t)]\hat{R}_D,$$

where  $D(t) = \text{diag}(B(t))$ ,  $E = Z^{-1}FZ$ , and  $\hat{R} = Z\hat{R}_D$ . Define

$$(3.3) \quad \text{cond}(Z) = \sup_{t \geq 0} \|Z(t)\| \cdot \|Z^{-1}(t)\|,$$

and assume that  $\text{cond}(Z)$  is bounded. In other words, we are assuming that  $Z$  is a Lyapunov transformation; e.g., this is certainly the case if the triangular system  $\dot{R} = BR$  is integrally separated. This last assertion follows from [11, Theorem 5.1].

Recall that  $R$  is an *integrally separated* fundamental matrix solution if there exist  $a > 0$  and  $d \geq 1$  such that

$$(3.4) \quad \frac{\|R(t)e_i\|}{\|R(s)e_i\|} \cdot \frac{\|R(s)e_{i+1}\|}{\|R(t)e_{i+1}\|} \geq de^{a(t-s)},$$

for all  $t, s : t \geq s \geq 0$ , and  $i = 1, 2, \dots, n-1$ . We also recall that integral separation is a generic property for linear systems ([19]) and is a necessary and sufficient condition for stability of the Lyapunov exponents, when they are distinct, [1].

Next, let  $\omega := \sup_{t \geq 0} \|E(t)\|$ , and observe that

$$(3.5) \quad \omega \leq \sup_{t \geq 0} \|F\| \operatorname{cond}(Z) \leq \delta \operatorname{cond}(Z).$$

We make note here that the integral separation constants used in Lemma 3.1 and Lemma 4.1 below are the integral separation constants for the piecewise constant upper triangular system that results from (2.4); see (2.9). That is, we write

$$R(t_{j+1}, t_j) = e^{(t_{j+1}-t_j)B(t)}, \quad t_j \leq t < t_{j+1},$$

where in fact

$$B(t) := \frac{1}{t_{j+1} - t_j} \log(R(t_{j+1}, t_j)), \quad t_j \leq t < t_{j+1}.$$

This piecewise constant triangular system produces the same upper triangular fundamental matrix solution as the exact upper triangular system when evaluated at mesh points. Therefore,

$$(3.6) \quad \int_{t_j}^{t_{j+1}} B_{ii}(\tau) d\tau = \int_{t_j}^{t_{j+1}} \tilde{B}_{ii}(\tau) d\tau$$

where  $B$  denotes the piecewise constant triangular coefficient matrix function and  $\tilde{B}$  the exact triangular coefficient matrix function of (1.3). Thus, if  $\tilde{B}$  has integral separation with constants  $\tilde{a} > 0$  and  $\tilde{d} \geq 0$  so that for  $t \geq s$  (take logarithms in (3.4))

$$(3.7) \quad \int_s^t (\tilde{B}_{ii}(\tau) - \tilde{B}_{i+1,i+1}(\tau)) d\tau \geq \tilde{a}(t-s) - \tilde{d}$$

for  $i = 1, \dots, n-1$ , then for  $t_{j-1} < s < t_j$  and  $t_k < t < t_{k+1}$ ,

$$(3.8) \quad \begin{aligned} & \int_s^t (B_{ii}(\tau) - B_{i+1,i+1}(\tau)) d\tau = \int_s^t (\tilde{B}_{ii}(\tau) - \tilde{B}_{i+1,i+1}(\tau)) d\tau \\ & + \int_s^{t_j} [(B_{ii}(\tau) - \tilde{B}_{ii}(\tau)) - (B_{i+1,i+1}(\tau) - \tilde{B}_{i+1,i+1}(\tau))] d\tau \\ & + \int_{t_k}^t [(B_{ii}(\tau) - \tilde{B}_{ii}(\tau)) - (B_{i+1,i+1}(\tau) - \tilde{B}_{i+1,i+1}(\tau))] d\tau \geq a(t-s) - d \end{aligned}$$

where  $a = \tilde{a}$ ,  $d \leq \tilde{d} + 4M_{i,i+1}h_{max}$ ,  $M_{i,i+1} = \sup_{t \geq 0} |B_{ii}(t) - B_{i+1,i+1}(t)|$ , and  $h_{max} = \sup_j (t_{j+1} - t_j)$ . In other words, also the problem with  $B$  has integral separation with constants  $a > 0$  and  $d \geq 0$ .

Then, the following result shows the existence of a near identity orthogonal change of variables which brings the perturbed diagonal system to upper triangular, provided that  $\omega$  is small enough. We proved this result in [14], under the assumption of integral separation of both unperturbed and perturbed triangular systems.

LEMMA 3.1. [14] Let  $\sup_{t \geq 0} \|B(t)\| = M$ , and let  $a$  and  $d$  be defined in (3.4). Let  $\hat{Q}$  be the orthogonal factor in the QR-factorization of  $\hat{R}_D$ .

If  $\omega < \omega_+(\alpha, K, M)$ , then  $|\hat{Q}_{ij}(t)| \leq \bar{\rho}$  for  $i \neq j$  and all  $t \geq 0$ . Here,  $\bar{\rho} = \beta \cdot \omega$ ,  $\beta = \alpha K$ ,  $\alpha > 1$ ,  $K = e^d/a$ , and

$$(3.9) \quad \omega_+(\alpha, K, M) := (\sqrt{a_1^2 + 4(\alpha-1)a_2 - a_1})/(2a_2)$$

where  $a_2 = n^2\beta^2[M\beta + 2]$  and  $a_1 = n\beta[2M\beta + 1]$ .

As an immediate consequence we have

**COROLLARY 3.1.** *If  $\omega < \omega_+(\alpha, K, M)$  and  $(n-1)\bar{\rho}^2 \leq 1$ , then  $\|\widehat{Q}(t) - I\| \leq \rho \equiv (n-1)(\bar{\rho} + \bar{\rho}^2)$  and  $\|\widehat{Q}(t) - I\|_F \leq \rho_F \equiv \bar{\rho}\sqrt{2(n^2 - n)}$ , where  $\|\cdot\|_F$  is the Frobenius norm.*

Perhaps surprisingly, we already have all the ingredients to obtain global error bounds on  $Q_k - Q(t_k)$ .

First of all, let us look again at (2.6). In the notation of Corollary 3.1, if  $\omega < \omega_+(\alpha, K, M)$ , then (2.6) can be rewritten as

$$(3.10) \quad X_k = Q(t_k)Z(t_k)\widehat{Q}(t_k)U(t_k)$$

where  $U(t_k)$  is upper triangular with positive diagonal elements, and  $\|\widehat{Q}(t_k) - I\| \leq \rho$ .

Next, let  $\widehat{Z} \equiv Z(t_k)$  and  $\widehat{Q} \equiv \widehat{Q}(t_k)$ . Then

$$(3.11) \quad \widehat{Z}\widehat{Q} = \widehat{Z}(I + \Delta\widehat{Q}) = \widehat{Z} + \widehat{Z}\Delta\widehat{Q} =: \widehat{Z} + \Delta\widehat{Z}$$

where  $\|\Delta\widehat{Z}\| \leq \|\widehat{Z}\| \cdot \|\Delta\widehat{Q}\| \leq \|\widehat{Z}\|\rho$ .

**THEOREM 3.2.** *With the previous notation, assume that*

1.  $\omega < \omega_+(\alpha, K, M)$ ,
2.  $\text{cond}(Z)\rho < 1/2$ , and
3.  $\|\widehat{Q}(t_k) - I\| \leq \rho$ .

*Then we have*

$$(3.12) \quad \|Q_k - Q(t_k)\| \leq \epsilon := \frac{3\text{cond}(Z)\rho}{1 - 2\text{cond}(Z)\rho}, \quad k = 0, 1, 2, \dots$$

*Proof.* By the perturbation theory for the  $QR$  factorization (e.g., see [22, Theorem 3.1]),

$$(3.13) \quad \widehat{Z} + \Delta\widehat{Z} = (I + W)(\widehat{Z} + G)$$

where  $I + W$  is orthogonal,  $\widehat{Z} + G$  is upper triangular with positive diagonal elements, and

$$(3.14) \quad \|W\| \leq \frac{3\|\widehat{Z}^{-1}\| \cdot \|\Delta\widehat{Z}\|}{1 - 2\|\widehat{Z}^{-1}\| \cdot \|\Delta\widehat{Z}\|}.$$

Thus, (3.10) may be written as

$$(3.15) \quad X_k = Q(t_k)(I + W)(\widehat{Z} + G)U(t_k)$$

and by the uniqueness of the  $QR$  factorization, from (2.5),  $Q_k = Q(t_k)(I + W)$  and therefore

$$(3.16) \quad \|Q_k - Q(t_k)\| = \|Q^T(t_k)Q_k - I\| = \|W\| \leq \epsilon.$$

□

It is possible to improve the perturbation bounds on the QR-factors of nearby matrices; e.g. [4, 5] and the references therein. However, the real drawback of the global error bound in (3.12) is actually due to the fact that we have used a global transformation (via  $Z$ ) to diagonal form, and are thus penalized by  $\text{cond}(Z)$ . The optimal situation of course is if  $Z = I$ , which occurs for instance when the upper triangular problem is in fact diagonal. In this case, we can take  $\omega = \delta$  in (3.5).

However, aside from this case of  $Z = I$ , it is probably best to avoid altogether the diagonalizing transformation  $Z$  and tackle directly the perturbed triangular problem in (3.1), thereby attempting to bring directly  $\widehat{R}$  to triangular form via an orthogonal near-the-identity transformation and obtain sharper estimates. This is what we do in the next Section.

**4. Handling the Triangular Term Directly.** Let us consider the perturbed triangular problem (3.1), rewritten here again as

$$(4.1) \quad \dot{\hat{R}} = [B(t) + F(t)]\hat{R}, \quad t \geq 0, \quad \hat{R}(0) = R_0,$$

with  $\sup_{t \geq 0} \|F(t)\| \leq \omega$ . Recall that  $B$  has upper triangular structure, and  $\omega$  is small. We have  $\omega = \delta$  here, see (2.10), but we chose to use  $\omega$  to unify the notation to Lemma 3.1.

Below, we show that there exists an orthogonal change of variables to upper triangular, that is a change of variables  $\hat{R} = \hat{Q}U$  with  $\hat{Q}$  orthogonal and  $U$  upper triangular with positive diagonal, such that  $\hat{Q}$  remains, under reasonable conditions, a small perturbation of the identity given the initial condition  $\hat{Q}(0) = \hat{Q}_0 = I$ . The proof of the Lemma below uses a similar technique to that used to prove Lemma 3.1 (see [14]), but much more careful estimates are now employed. In the simplest sense, Lemma 4.1 is a componentwise version of Lemma 3.1 but for a perturbed triangular system as opposed to a perturbed diagonal system. In order to obtain bounds on the entries of  $\hat{Q}$ , we assume bounds on the entries of  $B$  and assume integral separation constants for both consecutive and non-consecutive diagonal elements of  $B$ . Our bounds will be of the type  $|\hat{Q}_{ij}(t)| \leq \rho_{ij}$ ,  $i \neq j$ , with  $\rho_{ij} = \alpha_{ij}K_{ij}\omega$ ; see below. The key to obtaining this result is that the  $\alpha_{ij}$  in this bound may be found recursively starting from  $|i - j| = n - 1$  down to  $|i - j| = 1$  with, for instance,  $\alpha_{1,n} = \alpha_{n,1} = 2$ .

LEMMA 4.1. *Consider the problem (4.1), and write  $B(t) + F(t) = D(t) + T(t) + F(t)$ , for all  $t$ , where  $D = \text{diag}(B)$  and  $T = \text{upp}(B)$ . Also, let  $\sup_{t \geq 0} \|F(t)\| \leq \omega$ . Then, there exists an orthogonal change of variables  $\hat{Q}$ , with  $\hat{Q}(0) = I$ , which bring  $B + F$  to upper triangular structure  $C := \hat{Q}^T[B + F]\hat{Q} - \hat{Q}^T\dot{\hat{Q}}$ .*

*Moreover, let  $|D_{ii}(t)| \leq \kappa_{ii}$  for  $i = 1, \dots, n$ ,  $|T_{ij}(t)| \leq \kappa_{ij}$  for  $i < j$ , for all  $t \geq 0$ , and let  $K_{ij}$  be such that for all  $t \geq 0$ ,*

$$(4.2) \quad K_{ij} \geq \int_0^t e^{-\int_\tau^t (D_{ii}(r) - D_{jj}(r))dr} d\tau, \quad i < j \quad \text{and} \quad K_{ij} = K_{ji}, \quad i > j.$$

*For  $|i - j| = n - 1, |i - j| = n - 2, \dots, |i - j| = 1$ , choose  $\alpha_{ij}$  such that*

$$(4.3) \quad \alpha_{ij} > 1 + \sum_{k=j+1}^n \kappa_{jk}K_{ik}\alpha_{ik} + \sum_{k=1}^{i-1} K_{jk}\alpha_{jk}\kappa_{ki}, \quad \text{for } i < j, \quad \text{and let } \alpha_{ij} = \alpha_{ji}, \quad i > j.$$

*Set*

$$(4.4) \quad \omega_+^{(ij)} := (\sqrt{(a_1^{(ij)})^2 - 4a_0^{(ij)}a_2^{(ij)}} - a_1^{(ij)})/(2a_2^{(ij)}), \quad \omega_+ := \min_{i,j} \omega_+^{(ij)},$$

*where  $a_0^{(ij)}, a_1^{(ij)}, a_2^{(ij)}$  are defined in (4.17).*

*If  $\omega < \omega_+(\{\alpha_{ij}\}, \{K_{ij}\}, \{\kappa_{ij}\})$ , then  $|\hat{Q}_{ij}(t)| \leq \rho_{ij}$  for  $i \neq j$  and all  $t \geq 0$ , where  $\rho_{ij} = \alpha_{ij}K_{ij} \cdot \omega$ .*

*Proof.*

Recall, see (1.4), that  $\hat{Q}$  must satisfy  $\dot{\hat{Q}} = \hat{Q}S(\hat{Q}, B + F)$ . So, for  $i < j$  we have

$$(4.5) \quad \begin{aligned} \dot{\hat{Q}}_{ij} &= -\hat{Q}_{ij}[D_{ii} - D_{jj}] + \left( \hat{Q}_{ij}[D_{ii} - D_{jj}] + e_i^T(\hat{Q}[S(\hat{Q}, D) + S(\hat{Q}, T) + S(\hat{Q}, F)])e_j \right) \\ &=: -\hat{Q}_{ij}[D_{ii} - D_{jj}] + q_{ij}(t, \hat{Q}, \omega) \end{aligned}$$

and a similar formula for  $i > j$ . We want to show that if the conditions of the theorem are satisfied and  $\hat{Q}(0) = I$ , then  $|\hat{Q}_{ij}(t)| \leq \rho_{ij}$  for all  $i \neq j$  and  $t \geq 0$ . The proof involves applying [15, Theorem IV.2.1].

Using the nonlinear variation of constants formula we have for  $\hat{Q}(0) = I$  and  $i < j$ ,

$$(4.6) \quad \hat{Q}_{ij}(t) = \int_0^t e^{-\int_\tau^t (D_{ii}(r) - D_{jj}(r))dr} q_{ij}(\tau, \hat{Q}(\tau), \omega) d\tau$$

Thus,  $\sup_t |\hat{Q}_{ij}(t)| \leq K_{ij} \sup_t |q_{ij}(t, \hat{Q}(t), \omega)|$ . We have

$$(4.7) \quad |q_{ij}(t, \hat{Q}, \omega)| \leq |q_{ij}(t, \hat{Q}, \omega) - q_{ij}(t, I, \omega)| + |q_{ij}(t, I, \omega)| \leq \eta(\{\rho_{kl}\}, \omega)\rho_{ij} + N(\omega)$$

where since  $S(I, D) = S(I, T) = 0$  and  $S(I, F) = F_L - F_L^T$  where  $F_L = \text{low}(F)$ ,  $N(\omega) \leq \omega$ . To bound  $\eta(\{\rho_{kl}\}, \omega)$  write

$$(4.8) \quad \begin{aligned} q_{ij}(t, \hat{Q}, \omega) &= q_{ij}^D(t, \hat{Q}, \omega) + q_{ij}^T(t, \hat{Q}, \omega) + q_{ij}^F(t, \hat{Q}, \omega) \\ &:= \left( \hat{Q}_{ij}[D_{ii} - D_{jj}] + e_i^T \hat{Q} S(\hat{Q}, D) e_j \right) + e_i^T \hat{Q} S(\hat{Q}, T) e_j + e_i^T \hat{Q} S(\hat{Q}, F) e_j \end{aligned}$$

and consider the case in which  $i < j$  (the case  $i > j$  is similar).

For  $q_{ij}^D(t, \hat{Q}, \omega)$  we have from (A-3) using the notation  $\beta_{ij} = \alpha_{ij} K_{ij}$ ,

$$(4.9) \quad \begin{aligned} |q_{ij}^D(t, \hat{Q}, \omega) - q_{ij}^D(t, I, \omega)| &\leq \kappa_{ii}[\rho_{ij}(\rho_{ij}^2 + 2 \sum_{k=j+1}^n \rho_{ik}^2)] + \kappa_{jj}[\rho_{ij} \sum_{k=1, k \neq j}^n \rho_{jk}^2 + 2 \sum_{k=j+1}^n \rho_{ik} \rho_{jk}] \\ &+ \sum_{l \neq i, j} \kappa_{ll}[\rho_{lj}(\sum_{k=1}^{j-1} \rho_{ik} \rho_{lk} + \sum_{k=j+1}^n \rho_{ik} \rho_{lk})] \\ &\leq \kappa_{ii}[\rho_{ij}(\rho_{ij}^2 + 2 \sum_{k=j+1}^n \rho_{ik}^2)] + \kappa_{jj}[\rho_{ij} \sum_{k=1, k \neq j}^n \rho_{jk}^2 + 2 \sum_{k=j+1}^n \rho_{ik} \rho_{jk}] \\ &+ \sum_{l \neq i, j} \kappa_{ll}[\rho_{lj}(\rho_{li} + \rho_{il} + \sum_{k=1, k \neq i, j, l}^n \rho_{ik} \rho_{lk})] \\ &= \rho_{ij} \{ \omega^2 [\kappa_{ii}(\beta_{ij}^2 + 2 \sum_{k=j+1}^n \beta_{ik}^2) + \kappa_{jj} \sum_{k=1, k \neq j}^n \beta_{jk}^2 + \sum_{l \neq i, j} \kappa_{ll} \frac{\beta_{lj}}{\beta_{ij}} \sum_{k=1, k \neq i, j, l}^n \beta_{ik} \beta_{lk}] \\ &+ \omega [\frac{2\kappa_{jj}}{\beta_{ij}} \sum_{k=j+1}^n \beta_{ik} \beta_{jk} + \sum_{l \neq i, j} \kappa_{ll} \frac{\beta_{lj}}{\beta_{ij}} (\beta_{li} + \beta_{il})] \} =: \rho_{ij} \eta_{ij}^D =: \rho_{ij} (\omega^2 \eta_{ij}^{D,2} + \omega \eta_{ij}^{D,1}) . \end{aligned}$$

Next, we obtain using (A-4) and again  $\beta_{ij} = \alpha_{ij} K_{ij}$ :

$$(4.10) \quad \begin{aligned} |q_{ij}^T(t, \hat{Q}, \omega) - q_{ij}^T(t, I, \omega)| &\leq \sum_{m=j+1}^n \kappa_{jm} \rho_{im} + \sum_{l=1}^{i-1} \rho_{jl} \kappa_{li} + \sum_{l=1, l \neq j}^n \sum_{m=l+1, m \neq i}^n \rho_{jl} \kappa_{lm} \rho_{im} \\ &+ \sum_{k=1, k \neq i}^{j-1} \rho_{ik} \sum_{l=1}^n \sum_{m=l+1}^n \rho_{jl} \kappa_{lm} \rho_{km} + \sum_{k=j+1}^n \rho_{ik} \sum_{l=1}^n \sum_{m=l+1}^n \rho_{kl} \kappa_{lm} \rho_{jm} \\ &\leq \sum_{m=j+1}^n \kappa_{jm} \rho_{im} + \sum_{l=1}^{i-1} \rho_{jl} \kappa_{li} + \sum_{l=1, l \neq j}^n \sum_{m=l+1, m \neq i}^n \rho_{jl} \kappa_{lm} \rho_{im} \\ &+ \sum_{k=1, k \neq i}^{j-1} \rho_{ik} [ \sum_{m=j+1}^n \kappa_{jm} \rho_{km} + \sum_{l=1}^{k-1} \rho_{jl} \kappa_{lk} + \sum_{l=1, l \neq j}^n \sum_{m=l+1, m \neq k}^n \rho_{jl} \kappa_{lm} \rho_{km} ] \\ &+ \sum_{k=j+1}^n \rho_{ik} [ \sum_{m=k+1}^n \kappa_{km} \rho_{jm} + \sum_{l=1}^j \rho_{kl} \kappa_{lj} + \sum_{l=1, l \neq k}^n \sum_{m=l+1, m \neq j}^n \rho_{kl} \kappa_{lm} \rho_{jm} ] \\ &= \rho_{ij} \{ \omega^2 [ \sum_{k=1, k \neq i}^{j-1} \frac{\beta_{ik}}{\beta_{ij}} \sum_{l=1, l \neq j}^n \sum_{m=l+1, m \neq k}^n \beta_{jl} \beta_{km} \kappa_{lm} + \sum_{k=j+1}^n \frac{\beta_{ik}}{\beta_{ij}} \sum_{l=1, l \neq k}^n \sum_{m=l+1, m \neq j}^n \beta_{kl} \beta_{jm} \kappa_{lm} ] \\ &+ \omega [ \sum_{l=1, l \neq j}^n \sum_{m=l+1, m \neq i}^n \frac{\beta_{jl} \beta_{im} \kappa_{lm}}{\beta_{ij}} + \sum_{k=1, k \neq i}^{j-1} \frac{\beta_{ik}}{\beta_{ij}} ( \sum_{m=j+1}^n \kappa_{jm} \beta_{km} + \sum_{l=1}^{k-1} \beta_{jl} \kappa_{lk} ) \\ &+ \sum_{k=j+1}^n \frac{\beta_{ik}}{\beta_{ij}} ( \sum_{m=k+1}^n \kappa_{km} \beta_{jm} + \sum_{l=1}^{j-1} \beta_{kl} \kappa_{lj} ) ] + 1 [ \sum_{m=j+1}^n \kappa_{jm} \frac{\beta_{im}}{\beta_{ij}} + \sum_{l=1}^{i-1} \kappa_{li} \frac{\beta_{jl}}{\beta_{ij}} ] \} \\ &=: \rho_{ij} \eta_{ij}^T =: \rho_{ij} \eta_{ij}^{T,2} \omega^2 + \eta_{ij}^{T,1} \omega + \eta_{ij}^{T,0} . \end{aligned}$$

Then, using the fact that  $0 < \hat{Q}_{jj} \leq 1$ , and so  $1 - \hat{Q}_{jj} \leq 1 - \hat{Q}_{jj}^2 = \sum_{k=1, k \neq j}^n \hat{Q}_{jk}^2$ , we obtain from



(A-5),

$$\begin{aligned}
(4.11) \quad & |q_{ij}^F(t, \widehat{Q}, \omega) - q_{ij}^F(t, I, \omega)| \leq \omega \left[ \sum_{k \neq j, k=1}^n \rho_{jk}^2 + \sum_{k \neq i, k=1}^n \rho_{ik}^2 + \sum_{(l,m) \neq (j,i), l,m=1}^n \rho_{lj} \rho_{mi} + \sum_{k=1, k \neq i,j}^n \rho_{ik} \right] \\
& = \rho_{ij} \left\{ \omega^2 \left[ \sum_{k \neq j, k=1}^n \frac{\beta_{jk}^2}{\beta_{ij}} + \sum_{k \neq i, k=1}^n \frac{\beta_{ik}^2}{\beta_{ij}} + \sum_{(l,m) \neq (j,i), l,m=1}^n \frac{\beta_{lj} \beta_{mi}}{\beta_{ij}} \right] + \omega \sum_{k=1, k \neq i,j}^n \frac{\beta_{ik}}{\beta_{ij}} \right\} \\
& =: \rho_{ij} \eta_{ij}^F =: \rho_{ij} (\omega^2 \eta_{ij}^{F,2} + \omega \eta_{ij}^{F,1})
\end{aligned}$$

where  $\beta_{ij} = \alpha_{ij} K_{ij}$ .

So, we have using (4.7), (4.8), (4.9), (4.10), and (4.11),

$$(4.12) \quad \eta_{ij}(\{\rho_{kl}\}, \omega) \leq \eta_{ij}^D + \eta_{ij}^T + \eta_{ij}^F,$$

and finally from (4.7) we obtain

$$\sup_t |\widehat{Q}_{ij}(t)| \leq K_{ij} (N(\omega) + \eta_{ij}(\{\rho_{kl}\}, \omega) \rho_{ij}).$$

Since  $N(\omega) \leq \omega$ , Theorem IV.2.1 of [15] may be applied if

$$(4.13) \quad K_{ij}[(\eta_{ij}^D + \eta_{ij}^T + \eta_{ij}^F) \rho_{ij} + \omega] < \rho_{ij},$$

or

$$(4.14) \quad 1 > K_{ij}(\eta_{ij}^D + \eta_{ij}^T + \eta_{ij}^F) + 1/\alpha_{ij},$$

or equivalently

$$(4.15) \quad 0 > \alpha_{ij} K_{ij}(\eta_{ij}^D + \eta_{ij}^T + \eta_{ij}^F) + (1 - \alpha_{ij}).$$

Then we need to have

$$(4.16) \quad 0 > \alpha_{ij} K_{ij}(\eta_{ij}^{D,2} + \eta_{ij}^{T,2} + \eta_{ij}^{F,2}) \omega^2 + \alpha_{ij} K_{ij}(\eta_{ij}^{D,1} + \eta_{ij}^{T,1} + \eta_{ij}^{F,1}) \omega + [\alpha_{ij} K_{ij} \eta_{ij}^{T,0} + 1 - \alpha_{ij}],$$

which we rewrite as

$$(4.17) \quad f(\omega) := a_2^{(ij)} \omega^2 + a_1^{(ij)} \omega + a_0^{(ij)} < 0,$$

where in particular

$$(4.18) \quad a_0^{(ij)} = 1 - \alpha_{ij} + \sum_{l=1}^{i-1} \kappa_{li} \alpha_{jl} K_{jl} + \sum_{m=j+1}^n \kappa_{jm} \alpha_{im} K_{im}.$$

We notice that  $a_1^{(ij)} > 0$ ,  $a_2^{(ij)} > 0$ . Since  $f'(\omega) > 0$  for  $\omega > 0$ , we need to have  $a_0^{(ij)} < 0$  in order to be sure that there are values of  $\omega$  satisfying  $f(\omega) < 0$ . This is guaranteed by (4.3).

Thus, if  $\omega < \omega_+$  with  $\omega_+$  given in (3.9), then  $|\widehat{Q}_{ij}(t)| \leq \rho_{ij} \equiv \alpha_{ij} K_{ij} \omega$  for  $i \neq j$  and  $t \geq 0$ .

□

What one expects from Lemma 4.1 are  $\rho_{ij} \equiv \alpha_{ij} K_{ij} \omega$  where  $\rho_{ij}$  are smaller for  $|i - j|$  large than for  $|i - j|$  small, e.g.  $|i - j| = 1$ . This is due in part to the sharper bounds obtained by employing  $K_{ji} \equiv K_{ij} = e^{d_{ij}} / a_{ij}$  for  $i < j$  where for  $t \geq s$ ,

$$\int_s^t (B_{ii}(\tau) - B_{jj}(\tau)) d\tau \geq a_{ij}(t - s) - d_{ij},$$

as opposed to

$$\int_s^t (B_{ii}(\tau) - B_{jj}(\tau)) d\tau \geq \left[ \sum_{k=i}^{j-1} a_{k,k+1} \right] (t - s) - \sum_{k=i}^{j-1} d_{k,k+1},$$

essentially avoiding the use of a triangular inequality. It is also due to the form of the recursion to determine the  $\alpha_{ij}$  in (4.3) in which we choose  $\alpha_{1,n} = \alpha_{n,1} > 1$ , then determine  $\alpha_{ij}$  for  $|i-j| = n-2$ , etc.. There is the potential for the  $\alpha_{ij}$  to become large as  $|i-j| \downarrow 1$  depending on the size of the off diagonal elements of  $B$  characterized by  $\kappa_{ij}$  and the strength of the integral separation between diagonal elements as characterized by  $K_{ij}$ .

The bound on the perturbation  $F$  that allows for application of Lemma 4.1 is given in (4.4) using (4.17) with  $a_0^{(ij)}$  given in (4.18). The coefficients  $a_1^{(ij)}$ ,  $a_2^{(ij)}$  may be obtained from (4.9), (4.10), and (4.11) which give for  $\beta_{ij} := \alpha_{ij}K_{ij}$ ,

$$\begin{aligned}
 (4.19) \quad a_1^{(ij)} &:= \beta_{ij}(\eta_{ij}^{D,1} + \eta_{ij}^{T,1} + \eta_{ij}^{F,1}) \\
 &= [2\kappa_{jj} \sum_{k=j+1}^n \beta_{ik}\beta_{jk} + \sum_{l \neq i,j} \kappa_{ll}\beta_{lj}(\beta_{li} + \beta_{il})] + [\sum_{l=1, l \neq j}^n \sum_{m=l+1, m \neq i}^n \beta_{jl}\beta_{im}\kappa_{lm} \\
 &\quad + \sum_{k=1, k \neq i}^{j-1} \beta_{ik}(\sum_{m=j+1}^n \kappa_{jm}\beta_{km} + \sum_{l=1}^{k-1} \beta_{jl}\kappa_{lk})] + \sum_{k=j+1}^n \beta_{ik}(\sum_{m=k+1}^n \kappa_{km}\beta_{jm} + \sum_{l=1}^{j-1} \beta_{kl}\kappa_{lj}) + \sum_{k=1, k \neq i, j}^n \beta_{ik}
 \end{aligned}$$

and

$$\begin{aligned}
 (4.20) \quad a_2^{(ij)} &:= \beta_{ij}(\eta_{ij}^{D,2} + \eta_{ij}^{T,2} + \eta_{ij}^{F,2}) \\
 &= [\beta_{ij}\{\kappa_{ii}(\beta_{ij}^2 + 2 \sum_{k=j+1}^n \beta_{ik}^2) + \kappa_{jj} \sum_{k=1, k \neq j}^n \beta_{jk}^2\} + \sum_{l \neq i, j} \kappa_{ll}\beta_{lj} \sum_{k=1, k \neq i, j, l}^n \beta_{ik}\beta_{lk}] \\
 &\quad + [\sum_{k=1, k \neq i}^{j-1} \beta_{ik} \sum_{l=1, l \neq j}^n \sum_{m=l+1, m \neq k}^n \beta_{jl}\beta_{km}\kappa_{lm} + \sum_{k=j+1}^n \beta_{ik} \sum_{l=1, l \neq k}^n \sum_{m=l+1, m \neq j}^n \beta_{kl}\beta_{jm}\kappa_{lm}] \\
 &\quad + [\sum_{k \neq j, k=1}^n \beta_{jk}^2 + \sum_{k \neq i, k=1}^n \beta_{ik}^2 + \sum_{(l,m) \neq (j,i), l, m=1}^n \beta_{lj}\beta_{mi}].
 \end{aligned}$$

The following theorem show that  $C$  obtained from Lemma 4.1 is a small perturbation of the piecewise constant upper triangular  $B$  and that  $C$  may be interpreted as upper triangularizing a perturbation of  $A$  in which the perturbation is not in general small.

**THEOREM 4.2.** *Let  $\tilde{B}$  be the upper triangular matrix function obtained for the exact  $Q$ :  $\tilde{B} = Q^T A Q - S(Q, A)$ . If we write the piecewise constant upper triangular  $B$  as  $B = \tilde{B} + \tilde{F}$ , then the upper triangular  $C$  that results from having only approximated  $Q$  satisfies for  $\bar{Q} = Q\hat{Q}$ ,*

$$\begin{aligned}
 (4.21) \quad C &= \bar{Q}^T [A + G] \bar{Q} - S(\bar{Q}, A + G) \\
 &= \tilde{B} + \text{upp}[\bar{Q}^T (A + G) \bar{Q} - Q^T A Q] + (\text{low}[\bar{Q}^T (A + G) \bar{Q} - Q^T A Q])^T \\
 &= B + \text{upp}[-\tilde{F} + \bar{Q}^T (A + G) \bar{Q} - Q^T A Q] + (\text{low}[-\tilde{F} + \bar{Q}^T (A + G) \bar{Q} - Q^T A Q])^T
 \end{aligned}$$

where  $G = Q[\tilde{F} + F - S(Q, A)]Q^T$ .

Moreover, if Lemma 4.1 holds, then  $C$  is an  $O(\tilde{\rho})$  perturbation of  $B$  and  $C$  is an  $O(\tilde{\rho}) + O(\|\tilde{F}\|)$  perturbation of  $\tilde{B}$  for  $\tilde{\rho} = \max_{i \neq j} \rho_{ij}$ .

*Proof.* We have

$$\begin{aligned}
 (4.22) \quad C &= \hat{Q}^T [B + F] \hat{Q} - S(\hat{Q}, B + F) = \hat{Q}^T [\tilde{B} + \tilde{F} + F] \hat{Q} - S(\hat{Q}, \tilde{B} + \tilde{F} + F) \\
 &= \bar{Q}^T [A + G] \bar{Q} - S(\bar{Q}, A + G) = \text{upp}[\bar{Q}^T (A + G) \bar{Q}] + (\text{low}[\bar{Q}^T (A + G) \bar{Q}])^T \\
 &= \tilde{B} + \text{upp}[\bar{Q}^T (A + G) \bar{Q} - Q^T A Q] + (\text{low}[\bar{Q}^T (A + G) \bar{Q} - Q^T A Q])^T \\
 &= B - \tilde{F} + \text{upp}[\bar{Q}^T (A + G) \bar{Q} - Q^T A Q] + (\text{low}[\bar{Q}^T (A + G) \bar{Q} - Q^T A Q])^T \\
 &= B + \text{upp}[-\tilde{F} + \bar{Q}^T (A + G) \bar{Q} - Q^T A Q] + (\text{low}[-\tilde{F} + \bar{Q}^T (A + G) \bar{Q} - Q^T A Q])^T \\
 &= B + \text{upp}[-\tilde{F} + \bar{Q}^T (A + G) \bar{Q} - Q^T A Q] + (\text{low}[-\tilde{F} + \bar{Q}^T (A + G) \bar{Q} - Q^T A Q])^T
 \end{aligned}$$

where we have used that  $\text{upp}[\tilde{F}] = \tilde{F}$  and  $\text{low}[\tilde{F}] = 0$ . We now show that  $C$  is a small perturbation of  $B$ .

Since  $\overline{Q}^T G \overline{Q} = \widehat{Q}^T [\widetilde{F} + F] \widehat{Q} - \widehat{Q}^T S(Q, A) \widehat{Q}$ , we have

$$\widehat{Q}^T \widetilde{F} \widehat{Q} = \widetilde{F} + (\widehat{Q}^T - I) \widetilde{F} + \widetilde{F} (\widehat{Q} - I) + (\widehat{Q}^T - I) \widetilde{F} (\widehat{Q} - I),$$

and

$$\widehat{Q}^T S(Q, A) \widehat{Q} = S(Q, A) + (\widehat{Q}^T - I) S(Q, A) + S(Q, A) (\widehat{Q} - I) + (\widehat{Q}^T - I) S(Q, A) (\widehat{Q} - I).$$

Since  $\text{upp}[S(Q, A)] + (\text{low}[S(Q, A)])^T = 0$ , we then obtain (i) and (ii) follows from (4.22) and the definition of  $G$ .

□

Notice that although  $C$  results from having upper triangularized  $A + G$ , in general  $G$  is not small. We conclude this section with a few important remarks.

*Remarks 4.1.*

- (a) While Lemma 3.1 (see Lemma 3.1 of [14]) together with Theorem 3.2 of [14] are in a sense analogous to a classical Bauer-Fike Theorem by employing a diagonalizing transformation, Lemma 4.1 together with (4.21) may obtain sharper bounds by avoiding the use of a diagonalizing transformation.
- (b) We note here that if the original problem with coefficient matrix function  $A$  is integrally separated, then (see e.g. [1, 20, 21, 11, 12]) the  $\widetilde{B}$  problem is integrally separated, and hence the  $B$  problem is. Besides the bound on  $F$ ,  $\delta$ , and the measure of integral separation, the  $K_{ij}$ , the bounds obtained depend on the  $\kappa_{ij} = \sup_t |B_{ij}(t)|$  for  $i \leq j$ .
- (c) An important point to make here is that we can view the exact solution as a perturbation of the computed solution as opposed to the computed solution being a perturbation of the exact solution. The bound on the norm of the perturbation,  $\delta$ , is the same in either case, but by considering the exact solution as a perturbation of the computed solution, the quantities employed to bound the error, e.g.  $K_{ij}$  and  $\kappa_{ij}$ , may be obtained from the computed solution.

**4.1. Simplified Bounds and Approximations.** Next, we derive somewhat simplified bounds on  $\|\widehat{Q}(t) - I\|$  by first taking the largest of the  $\rho_{ij}$ , then using  $\rho_k = \sup_{k=|i-j|} \rho_{ij}$  and  $\kappa = \sup_{i,j} \kappa_{ij} = \sup_{i,j} |B_{ij}(t)|$ . In addition, we determine an asymptotic approximation for  $\omega_+$ . Note that we have  $\rho_{ij} = \rho_{ji}$  for  $i \neq j$ , so the bounds we obtain on  $\|\widehat{Q}(t) - I\|$  are identical in the 1, 2, and  $\infty$  norms.

**COROLLARY 4.1.** *In the notation of Lemma 4.1, let  $\delta \equiv \omega < \omega_+(\{\alpha_{ij}\}, \{K_{ij}\}, \{\kappa_{ij}\})$ . Let  $\widetilde{\rho} = \max_{i \neq j} \rho_{ij}$ , and assume that  $(n-1)(\widetilde{\rho} + \widetilde{\rho}^2) \leq 1$ . Then,  $\|\widehat{Q}(t) - I\| \leq \rho \equiv (n-1)(\widetilde{\rho} + \widetilde{\rho}^2)$  and  $\|\widehat{Q}(t) - I\|_F \leq \rho_F \equiv \sqrt{2(n^2 - n)} \widetilde{\rho}$ . for all  $t \geq 0$ . Moreover, for  $k = 0, 1, 2, \dots$ ,  $\|Q_k - Q(t_k)\| \leq \rho$ .*

We next prove a corollary that gives a more computable bound on the error in the approximate  $Q$  while taking into account the variation in  $\rho_{ij}$  as a function of  $i$  and  $j$ . Let  $\rho_k = \sup_{k=|i-j|} \rho_{ij}$ ,  $K_k = \sup_{k=|i-j|} K_{ij}$ , and  $\alpha_k = \sup_{k=|i-j|} \alpha_{ij}$ , and assume that  $\kappa = \sup_{i,j} \kappa_{ij}$ .

**COROLLARY 4.2.** *If the assumptions of Lemma 4.1 are satisfied,  $K_{n-1} \leq K_{n-2} \leq \dots \leq K_1$ , and  $\sum_{k=1}^{n-1} \rho_k + \rho_k^2 \leq 1$ , then  $\rho_j \leq \alpha_j K_j \omega$ ,  $\|\widehat{Q}(t) - I\|_F \leq (2 \sum_{k=1}^{n-1} (n-k) \rho_k^2 + \sum_{i=1}^n (\sum_{k=1}^{i-1} \rho_k^2 + \sum_{i=1}^{n-i} \rho_k^2)^2)^{1/2}$ , and*

$$(4.23) \quad \|\widehat{Q}(t) - I\| \leq \begin{cases} 2 \sum_{k=1}^{(n-1)/2} (\rho_k + \rho_k^2), & n \text{ odd}, \\ 2 \sum_{k=1}^{(n-2)/2} (\rho_k + \rho_k^2) + (\rho_{n/2} + \rho_{n/2}^2), & n \text{ even}, \end{cases}$$

where  $\alpha_{n-1} > 1$  and

$$(4.24) \quad \alpha_j > 1 + (n-j-1)\kappa K_{j+1}(1 + (n-j-2)\kappa K_{j+2}(1 + \dots(1 + \kappa K_{n-1}\alpha_{n-1}))) \dots$$

for  $j = n-2, \dots, 1$ .

*Proof.* The proof involves recursively applying the condition (4.3) for  $i < j$  and  $k = |i-j| = n-1, \dots, 1$ . This means we can choose  $\alpha_k$  so that

$$(4.25) \quad \alpha_{n-1} > 1, \quad \alpha_{n-2} > 1 + \kappa K_{n-1} \alpha_{n-1} > \alpha_{n-1}, \quad \alpha_{n-3} > 1 + 2\kappa K_{n-2} \alpha_{n-2} > \alpha_{n-2}, \dots$$

or in general

$$(4.26) \quad \alpha_{j-1} > 1 + (n-j)\kappa K_j \alpha_j > \alpha_j, \quad j = n-1, \dots, 1.$$

The result then follows by recalling that  $\rho_j \leq \alpha_j K_j \omega$  and obtaining the bounds on the matrix norms in a straightforward fashion using  $\|\widehat{Q}(t) - I\| \leq \sqrt{\|\widehat{Q}(t) - I\|_1 \|\widehat{Q}(t) - I\|_\infty}$ .  $\square$

In spite of their appearances, Corollaries 4.1 and 3.1 are very different, even if in Corollary 3.1 we happen to have  $\omega = \delta$  (as when  $Z = I$ ). Moreover, the factors  $\omega_+$ , as well as  $\rho$ , are different in these two contexts; notice the use of  $\bar{\rho}$  in Corollary 3.1 versus the use of  $\tilde{\rho}$  in Corollary 4.1.

To estimate  $\omega_+$ , we reason as follows. An asymptotic analysis of the terms in (4.20), (4.19), and (4.18) that contribute to  $a_2, a_1$ , and  $a_0$  in (4.16) and (4.17) which determine  $\omega_+$  in (4.4) suggest that

$$(4.27) \quad a_2 \approx (3\kappa_D + \kappa)\alpha_1^3 K_1^3, \quad a_1 \approx (2\kappa_D + \kappa)\alpha_1^2 K_1^2, \quad \text{and} \quad a_0 \approx 1 - \alpha_1 + 2\kappa\alpha_2 K_2 < 0$$

where  $\kappa_D = \sup_{i,t} |B_{ii}(t)|$ .

These coefficients were determined by first observing that the dominant term for  $a_2 \equiv \beta_{ij}(\eta_{ij}^{D,2} + \eta_{ij}^{T,2} + \eta_{ij}^{F,2})$  in (4.20) where  $\beta_{ij} = \alpha_{ij} K_{ij}$  is proportional to  $\alpha_1^3 K_1^3$ . Then in (4.20) we obtain the term  $3\kappa_D$  from  $\kappa_{ii}\beta_{ij}^3 + \kappa_{jj}\beta_{ij}(\beta_{j,j-1} + \beta_{j,j+1})$  when  $j = i+1$  and the term  $\kappa$  is obtained from  $\beta_{ik}\beta_{jl}\beta_{km}\kappa_{lm}$  when  $k = i+1, l = j-1, m = i+2$  and  $j = i+2$ . There are no terms in  $\eta_{ij}^{F,2}$  proportional to  $\alpha_1^3 K_1^3$ . To determine the dominant terms in  $a_1$  observe that the term  $2\kappa_D$  is obtained from  $\kappa_{ll}\beta_{lj}(\beta_{li} + \beta_{il})$  in (4.19) when  $l = j-1 = i+1$  and the term  $\kappa$  is obtained from  $\beta_{jl}\beta_{im}\kappa_{lm}$  in (4.19) when  $l = j-1, m = i+1$ , and  $j = i+1$ . The dominant terms in  $\eta_{ij}^{F,1}$  are not proportional to  $\alpha_1^2 K_1^2$ . The approximation of  $a_0$  is found by considering (4.18) when  $j = i+1$ .

Using the form for  $\omega_+$  in (4.4) and the approximation  $\sqrt{1+x} \approx 1 + \frac{x}{2}$ , we have  $\omega_+ \approx -\frac{a_0}{a_1} \approx C[(2\kappa_D + \kappa)\alpha_1^2 K_1^2]^{-1}$  where  $C \approx -a_0$ .

Notice that the  $\rho_{ij}$  we have when treating the triangular term directly appears to decrease as  $|i-j|$  grows. There is an accumulation for  $|i-j|$  small, but it looks friendlier than the accumulation to find  $\text{cond}(Z)$  when using the diagonalizing transformation. Indeed, it is interesting to compare the bounds one obtain with the two different approaches: (i) using the diagonalizing transformation  $Z$ , and (ii) working directly with the triangular system. We do this here below on a two-dimensional system.

**4.2. Comparison for Two-Dimensional Systems.** Here we compare the two global error bounds obtained by the two different approaches we examined: (i) using the diagonalizing transformation  $Z$ , see Section 3, and (ii) dealing directly with the triangular coefficient matrix function (call this the *triangular approach*), for a two-dimensional system. That is, we have  $B = \begin{pmatrix} D_{11} & T_{12} \\ 0 & D_{22} \end{pmatrix}$ .

Of course, a bound on the error in  $Q$  using the diagonalizing transformation approach is given by (3.12), while with the triangular approach it is given in Corollary 4.1. The interesting thing is to see what bounds we need for  $\omega$  in the two cases.

We assume  $|D_{ii}(t)| \leq M$  for  $i = 1, 2$  and  $|T_{12}(t)| \leq \kappa_{12}$  for all  $t$  and use the bound  $\|Z(t)\|_F, \|Z^{-1}(t)\|_F \leq \sqrt{2 + \kappa_{12}^2 K^2}$  where  $K = e^d/a$  (see Lemma 3.1).

Then, we have a bound  $|\sin(\theta(t))| \leq \rho := 2K\omega$  where: (i)  $|\sin(\theta(t))| = |\widehat{Q}_{12}(t)|$  and  $\omega = \delta \text{cond}(Z)$  with the diagonalizing transformation approach, and (ii)  $|\sin(\theta(t))| = |\widehat{Q}_{12}(t)|$  and  $\omega = \delta$  with the triangular approach, provided that in these two cases we have

$$(4.28) \quad \text{(i)} \quad \omega < \omega_+^D := \frac{\sqrt{1+4MK} - 1}{2K(2MK+1)} \cdot \frac{1}{\text{cond}(Z)}$$

and

$$(4.29) \quad \text{(ii)} \quad \omega < \omega_+^T \equiv \frac{\sqrt{1+4M/(K\kappa_{12}^2)} - 1}{8MK/\kappa_{12}},$$

respectively.

Since for  $x > 0$ ,  $1 + x/2 - x^2/8 \leq \sqrt{1+x} \leq 1 + x/2$ , using  $\text{cond}(Z) \leq 2 + K^2\kappa_{12}^2$ , we have

$$(4.30) \quad \frac{M(1 - MK)}{(2MK + 1)\text{cond}(Z)} < \omega_+^D < \frac{M}{(2MK + 1)\text{cond}(Z)} < \frac{1}{2K\text{cond}(Z)}$$

and

$$(4.31) \quad \frac{1}{4K^2\kappa_{12}} \left[ 1 - \frac{M}{K\kappa_{12}^2} \right] = \frac{1 - M/(K\kappa_{12}^2)}{4K^2\kappa_{12}} < \omega_+^T < \frac{M/(K\kappa_{12}^2)}{4MK/\kappa_{12}} = \frac{1}{4K^2\kappa_{12}} \leq \frac{1}{4K\sqrt{\text{cond}(Z) - 2}}.$$

Quite clearly, the triangular approach gives much improved bounds.

**4.3. Bounding the  $K_{ij}$ 's.** In light of Remark 4.1(c), besides the size of the perturbation of  $B$ , the quantities needed to apply Lemma 4.1, the  $\kappa_{ij}$ 's that measure the non-normality and the  $K_{ij}$ 's that measure the integral separation, can be obtained from the computed solution. To this end we consider now how to obtain bounds for the  $K_{ij}$ 's. We consider two approaches.

The first approach follows ideas developed in [11] using Steklov averages, see also Adrianova [1] Lemma 5.4.1. In particular, for  $i < j$  set

$$p(t) := B_{ii}(t) - B_{jj}(t)$$

and for some  $H > 0$ , consider finding the quantities  $c(H)$  (positive) and  $M$ :

$$(4.32) \quad c(H) = \inf_t \frac{1}{H} \int_t^{t+H} p(r) dr > 0, \quad M = \inf_t p(t).$$

We have the following result.

LEMMA 4.3. *For  $a := c(H) > 0$ , and a positive integer  $N$ , let*

$$d := \left\{ \sum_{k=1}^N (c(H) - c(H/2^k)) \cdot (H/2^k) \right\} + (c(H) - M) \cdot (H/2^N) \geq 0.$$

*Then, for  $t \geq s$ , we have*

$$(4.33) \quad \int_s^t p(r) dr \geq a(t - s) - d.$$

*Proof.* That  $d \geq 0$  follows from

$$c(H) \geq c(H/2) \geq c(H/4) \geq \dots \geq c(H/2^N) \geq M.$$

If  $t - s = jH$  for some positive integer  $j$ , then

$$\int_s^t p(r) dr \geq c(H) \cdot (t - s) \geq a(t - s) - d$$

since  $d \geq 0$ . Otherwise,  $t - s = jH + (t - s - jH)$  with  $0 < (t - s - jH) < H$  and for  $\gamma_k \in \{0, 1\}$  we can write

$$t - s - jH = \left\{ H \sum_{k=1}^N \gamma_k / 2^k \right\} + x$$

where  $0 \leq x < H/2^N$ .

$$(4.34) \quad \begin{aligned} \int_s^t p(r) dr &\geq jH \cdot c(H) + \left\{ H \sum_{k=1}^N \gamma_k 2^k \cdot c(H/2^k) \right\} + x \cdot M \\ &= c(H) \cdot (t - s) + (jH - (t - s))c(H) + \left\{ H \sum_{k=1}^N \gamma_k / 2^k \cdot c(H/2^k) \right\} + x \cdot M \\ &\geq a(t - s) - d. \end{aligned}$$

□

Recalling (4.2), based on the above Lemma, we may use

$$K_{ij} = e^d/a.$$

Of course, the  $K_{ij}$ 's are still functions of  $H$ :  $K_{ij}(H)$ . The idea now is to use for  $K_{ij}$  the minimum value of  $K_{ij}(H)$  subject to maintaining  $c(H) > 0$ .

We also develop an alternative approach that is a simplification of Lemmas 4.1 and 4.2 of [17]. This alternative approach may yield better bounds on the  $K_{ij}$ 's.

As before, for all  $t$ , let  $p(t) = B_{ii}(t) - B_{jj}(t)$ ,  $i < j$ . Consider a discretization of the interval  $[0, T]$ :

$$0 = t_0 < t_1 < \dots < t_N = T.$$

LEMMA 4.4. *Let  $\epsilon > 0$  be given. There exists  $a_k > 0$  and  $d_k \geq 0$  such that for  $t_k \leq s \leq t_{k+1}$ ,*

$$(4.35) \quad \int_s^{t_{k+1}} p(r) dr \geq a_k(t_{k+1} - s) - d_k.$$

where for  $h_k = t_{k+1} - t_k$  and for

$$(4.36) \quad Y_k = \min_{t_k \leq s \leq t_{k+1}} \frac{1}{t_{k+1} - s} \int_s^{t_{k+1}} p(r) dr,$$

$$(4.37) \quad a_k = \begin{cases} \epsilon h_k^{-1}, & h_k Y_k < \epsilon, \\ Y_k, & h_k Y_k \geq \epsilon, \end{cases} \quad d_k = \begin{cases} \epsilon - h_k Y_k, & h_k Y_k < \epsilon, \\ 0, & h_k Y_k \geq \epsilon, \end{cases}$$

and

$$(4.38) \quad \int_0^T e^{-\int_s^T p(r) dr} ds \leq \sum_{k=0, d_k=0}^{N-1} e^{-\left[\sum_{l=k+1}^{N-1} X_l\right]} \frac{(1 - e^{-Y_k h_k})}{Y_k} + \frac{(e^\epsilon - 1)}{\epsilon} \cdot \sum_{k=0, a_k=\epsilon h_k^{-1}}^{N-1} e^{-\left[\sum_{l=k+1}^{N-1} X_l\right]} h_k e^{-h_k Y_k} =: K(T)$$

where

$$X_k = \int_{t_k}^{t_{k+1}} p(r) dr.$$

*Proof.* If  $h_k Y_k < \epsilon$ , then for  $a_k = \epsilon/h_k$ ,  $d_k = \epsilon - h_k Y_k \geq 0$ , and  $s \in [t_k, t_{k+1}]$ ,

$$(4.39) \quad \int_s^{t_{k+1}} p(r) dr \geq (t_{k+1} - s) Y_k = (t_{k+1} - s) \left[ \frac{\epsilon}{h_k} - \frac{(\epsilon - h_k Y_k)}{h_k} \right] \geq \frac{\epsilon}{h_k} (t_{k+1} - s) - (\epsilon - h_k Y_k).$$

If  $h_k Y_k \geq \epsilon$ , then for  $a_k = Y_k$ ,  $d_k = 0$ , and  $s \in [t_k, t_{k+1}]$ ,

$$(4.40) \quad \int_s^{t_{k+1}} p(r) dr \geq (t_{k+1} - s) Y_k = a_k(t_{k+1} - s).$$

The proof of (4.38) is then a direct consequence of the following estimate

$$\begin{aligned} \int_0^T e^{-\int_s^T p(r) dr} ds &= \sum_{k=0}^{N-1} \int_{t_k}^{t_{k+1}} e^{-\int_s^T p(r) dr} ds \\ &= \sum_{k=0}^{N-1} \int_{t_k}^{t_{k+1}} e^{-\left[\int_s^{t_{k+1}} p(r) dr + \sum_{l=k+1}^{N-1} \int_{t_l}^{t_{l+1}} p(r) dr\right]} ds \\ &\leq \sum_{k=0}^{N-1} e^{-\left[\sum_{l=k+1}^{N-1} X_l\right]} \int_{t_k}^{t_{k+1}} e^{-a_k(t_{k+1}-s)+d_k} ds = \sum_{k=0}^{N-1} e^{-\left[\sum_{l=k+1}^{N-1} X_l\right]} \frac{e^{d_k}}{a_k} (1 - e^{-a_k h_k}). \end{aligned}$$

□

Observe that the bounds (4.38) can be used to obtain bounds on the  $K_{ij}$ 's in (4.2) by setting  $K_{ij} = \sup_k K(t_k)$ .

**5. Example.** Here we illustrate our results, in particular the goodness of the bounds on the error in the orthogonal matrix function  $Q$ , on the following example, also considered in [14]. We only report on the improved bounds obtained when handling directly the triangular term.

Let  $B(t) = D(t) + U(t)$  be the upper triangular matrix function with

$$(5.1) \quad D(t) = \text{diag}(D_{11}(t), D_{22}(t), D_{33}(t), D_{44}(t)),$$

where we take  $D_{11}(t) = 10 + \sin(t)$ ,  $D_{22}(t) = \zeta \cos(t)$ ,  $D_{33}(t) = \lambda - \zeta \cos(t)$ ,  $D_{44}(t) = -10 + \sin(t)$ ,  $\zeta > 0$ , and

$$(5.2) \quad U(t) = \kappa \begin{pmatrix} 0 & \cos(t) & \sin(t) & \cos(t) \\ 0 & 0 & \cos(t) & \sin(t) \\ 0 & 0 & 0 & \cos(t) \\ 0 & 0 & 0 & 0 \end{pmatrix}.$$

The parameter  $\kappa$  changes the degree to which there is non-normality in the upper triangular part and the parameters  $\lambda$  and  $\zeta$  determine the degree of integral separation in the system. For simplicity, in our experiments below we fix  $\lambda = -5$ , and all computations refer to this case.

Form the matrix function

$$A(t) = Q(t)B(t)Q^T(t) + \dot{Q}(t)Q^T(t),$$

where

$$Q(t) = \text{diag}(1, Q_\beta(t), 1) \cdot \text{diag}(Q_\eta(t), Q_\eta(t)),$$

and

$$Q_\gamma(t) = \begin{pmatrix} \cos(\gamma t) & \sin(\gamma t) \\ -\sin(\gamma t) & \cos(\gamma t) \end{pmatrix}, \quad \eta = 1, \quad \beta = \sqrt{2}.$$

Results for this problem were obtained using the code LESLIS, which we developed

(see [www.math.gatech.edu/~dieci](http://www.math.gatech.edu/~dieci) and [www.math.ku.edu/~evanvleck](http://www.math.ku.edu/~evanvleck)).

In particular, we employ the continuous  $QR$  method (**ContQR** in Table 5.1) using the projected 5th order scheme (**IPAR(8)=0** in **LESLIS**) with local error control on the orthogonal factor  $Q$  (**IPAR(10)=1** in **LESLIS**), and the discrete  $QR$  method (**DiscQR** in Table 5.1) with a 5th order scheme (**IPAR(8)=4** in **LESLIS**), with local error control on the Lyapunov exponents (**IPAR(10)=0** in **LESLIS**). **TOL** is the value of the local error tolerance, and we used  $\text{TOL} = 1.E - 6$  throughout. Before reporting on the results, we remark that, see (2.11), we expect to have  $\omega$  to be about the same as **TOL**. In other words, the bound on the norm of the perturbation term  $F$  in Lemma 4.1 is essentially **TOL**. In Table 5.1 we tabulate the actual error for different methods, and  $\kappa$  and  $\zeta$  values. We report on the error in  $Q$  in the two norm (the largest singular value of the error) at gridpoints. Exponential notation is used throughout.

We further compare the actual error with the error bounds obtained in the previous sections. Although the quantities needed to determine the bounds,  $\rho_{ij} = \alpha_{ij}K_{ij}\omega$ , on  $|\hat{Q}_{ij}(t)|$ , in particular the  $\alpha_{ij}$ , are somewhat difficult to give in closed form, the recursion (4.3) are straightforward to code. Likewise, the formula for  $\omega_+$ , given by (4.4) with (4.18), (4.19), and (4.20) are functions of the  $K_{ij}$ , the measure of integral separation, and the  $\kappa_{ij}$ , the measure of non-normality. For this problem, we have  $\kappa_{11} = \kappa_{44} = 11$ ,  $\kappa_{22} = \zeta$ ,  $\kappa_{33} = |\lambda| + \zeta$ , and  $\kappa_{ij} = \kappa$  for  $i < j$ .

We have, for  $\lambda = -5$  and  $\zeta < \sqrt{24}$ ,  $K_{12} = K_{24} = 1/(10 - \sqrt{1 + \zeta^2})$ ,  $K_{13} = 1/(15 - \sqrt{1 + \zeta^2})$ ,  $K_{34} = 1/(5 - \sqrt{1 + \zeta^2})$ , and  $K_{14} = 1/20$ . If  $\zeta < 5/2$ , then  $K_{23} = 1/(5 - 2\zeta)$ . We next focus on determining bounds on  $K_{23}$  using Lemma 4.3 and Lemma 4.4 when  $\zeta \geq 5/2$ .

We have  $p(t) := D_{22}(t) - D_{33}(t) = 5 + 2\zeta \cos(t)$ . To employ Lemma 4.3 we have

$$(5.3) \quad K_{23}(H) = \frac{e^{d(H)}}{c(H)}, \quad d(H) = 4\zeta \left\{ H/2^{N+1} + \sum_{k=1}^N \sin(H/2^{k+1}) - \sin(H/2) \right\}$$

$$T = 10^4, \lambda = -5, \text{TOL}=1.E-6.$$

$\kappa$	$\zeta$	Method	Error	Method	Error
0	1	Cont QR	6E-7	Disc QR	7E-8
0	2	Cont QR	6E-7	Disc QR	1E-7
0	4	Cont QR	3E-6	Disc QR	5E-7
1	1	Cont QR	6E-7	Disc QR	7E-8
1	2	Cont QR	6E-7	Disc QR	1E-7
1	4	Cont QR	2E-6	Disc QR	5E-7
10	1	Cont QR	6E-7	Disc QR	1E-7
10	2	Cont QR	6E-7	Disc QR	2E-7
10	4	Cont QR	9E-6	Disc QR	5E-7
100	1	Cont QR	5E-6	Disc QR	4E-7
100	2	Cont QR	7E-6	Disc QR	4E-7
100	4	Cont QR	1E-4	Disc QR	2E-6

TABLE 5.1

Error in the approximate  $Q$  varying the degree of non-normality and integral separation, method and tolerance.

$$K_{23} \text{ bounds for } T = 10^4, \lambda = -5, \text{ and } \zeta = 4.$$

$N$	$K_{23}$ w/ Lemma 4.3	$K_{23}$ w/ Lemma 4.4 ( $h = \pi/2^N, \epsilon = 10^{-8}$ )
0	2.1E5	3.9E4
1	4.7E4	2.1E2
2	3.9E4	9.0E1
3	3.8E4	5.1E1
4	3.8E4	4.1E1
5	3.8E4	3.9E1
6	3.8E4	3.8E1

TABLE 5.2

Bounds on  $K_{23}$  obtained using Lemmas 4.3 and 4.4.

for  $H$  sufficiently large so that  $c(H) \equiv 5 - 4\zeta \sin(H/2)/H > 0$ . For different values of  $N$  we can determine, at least approximately, optimal values of  $H$ . When using Lemma 4.4, and in the notation used there, we set  $h \equiv h_k = \pi/2^N$  for integer  $N \geq 0$ . When  $p(t)$  is decreasing,  $Y_k = 5 + 2\zeta \cos(t_{k+1})$ , and when  $p(t)$  is increasing,  $Y_k = 5 + 2\zeta [\sin(t_{k+1}) - \sin(t_k)]/h$ .

In Table 5.2 we tabulate bounds on  $K_{23}$  found using Lemmas 4.3 and 4.4 for different values of  $N$ . In the case of Lemma 4.3  $N$  refers to discretization of the Steklov window length as in (5.3), while in the case of Lemma 4.4,  $N$  refers to the fineness of the discretization of  $[0, T]$  ( $h = \pi/2^N$ ). We will use the best of the bounds found to determine bounds on the error in  $Q$ .

By Corollary 4.2 we have  $\|Q(t_k) - Q_k\| \approx 2\rho_1 + \rho_2$ , and we recall that  $\rho_i \leq \alpha_i K_i \omega$ ,  $i = 1, 2$ . For convenience, in Table 5.3 we report on the estimates for  $\rho_1$  and  $\rho_2$  obtained by using these bounds and those for  $\alpha_i, K_i$ ,  $i = 1, 2$ , for different values of  $\kappa$  and  $\zeta$ .

At this point, we make an important observation: *If we use  $\omega = \text{TOL}$  in the expressions obtained for  $\rho_1$  and  $\rho_2$  in Table 5.3, and use  $2\rho_1 + \rho_2$  as a measure of the true global error, then we observe that the observed true global error is always below the theoretical estimate.* This is actually an important fact, since it is not easy to know exactly the value of  $\omega$ . At the same time, this is not entirely rigorous, since we can use  $2\rho_1 + \rho_2$  as a measure of the global error, only if  $\omega < \omega_+$ . For this reason, we also estimated  $\omega_+$  using the approximations to  $a_2, a_1, a_0$  in (4.27).

For the example we are considering,  $\kappa_D := \sup_{i,t} |B_{ii}(t)| = 11$ . In the last column of Table 5.3 we record this estimated value of  $[(2\kappa_D + \kappa)\alpha_1^2 K_1^2]^{-1}$  which serves as the estimate of  $\omega_+$  (see the discussion after (4.27)).



Bounds obtained for the example with  $\lambda = -5$ .

$\kappa$	$\zeta$	$\rho_1/\omega$	$\rho_2/\omega$	$(2\rho_1 + \rho_2)/\omega$	$\omega_+ \approx [(2\kappa_D + \kappa)\alpha_1^2 K_1^2]^{-1}$
0	1	6.6E-1	2.3E-1	1.6E0	8.5E-2
0	2	2.0E0	2.6E-1	4.3E0	2.6E-2
0	4	7.4E1	3.4E-1	1.5E2	2.7E-4
1	1	8.0E-1	2.4E-1	1.9E0	6.5E-2
1	2	2.4E0	2.7E-1	5.1E0	1.9E-2
1	4	9.4E1	3.6E-1	1.9E2	1.1E-4
10	1	2.6E0	3.5E-1	5.6E0	8.3E-3
10	2	8.2E0	3.9E-1	1.7E1	1.3E-3
10	4	3.6E2	5.1E-1	7.2E2	7.6E-7
10 <sup>2</sup>	1	7.7E1	1.4E0	1.6E2	1.6E-6
10 <sup>2</sup>	2	2.5E2	1.5E0	5.0E2	1.6E-7
10 <sup>2</sup>	4	1.2E4	2.0E0	2.4E4	7.4E-11

TABLE 5.3

Bound on the error in  $Q$  is dominated by  $\rho_1 = \alpha_1 K_1 \omega$ .

Now, by comparison of the true errors obtained in Table 5.1, using the estimates for  $\omega_+$  from Table 5.3, and still adopting the rationale that  $\omega = \text{TOL}$ , we see that the results for  $\kappa = 0$ ,  $\kappa = 1$ , or  $\kappa = 10$  and  $\zeta \neq 4$  in Table 5.3 validate that  $\omega < \omega_+$ . To further corroborate this fact, we notice that, for these value of  $\kappa$ , the error bounds obtained are in good agreement with the actual errors observed and recorded in Tables 5.1 and 5.3. For  $\kappa = 10^2$  or  $\kappa = 10$  and  $\zeta = 4$  the bounds obtained do not appear to be sharp, in the sense that it becomes increasingly difficult to satisfy the hypothesis  $\omega < \omega^+$  that is necessary to apply Lemma 4.1 and Corollary 4.2.

**6. Conclusions and Consequences.** We have provided a global error analysis for the factor  $Q$  in the change of variables  $X = QR$  of a fundamental matrix solution for a non-autonomous linear system. Our basic technique consists of a combination of backward and forward error analyses, and this is seemingly a new approach in the context of numerical integration.

Several comments are in order.

- (1) Among the noteworthy consequences of a global error analysis for  $Q$ , probably the most important one is that it becomes simple to obtain global error statements for all quantities which derive from the simplified triangular structure of the linear system. For example, error bounds on the Lyapunov exponents and/or the Sacker-Sell spectrum.
- (2) We have made our analysis for the entire fundamental matrix solution, although at times one is interested in computing the QR factorization of only a few of the columns of the fundamental matrix solution: *reduced* QR factorization. But, to deal with this case is—in principle—simple, since as a consequence of the combined backward and forward error analyses for all  $k$  we have

$$\begin{aligned}
 Q_k &= Q(t_k)\tilde{Q}(t_k), \\
 (6.1) \quad R_k R_{k-1} \dots R_2 R_1 R(t_0) &= \tilde{Q}^T(t_k)[R(t_k, t_{k-1}) + E_k] \dots [R(t_2, t_1) + E_2][R(t_1, t_0) + E_1]R(t_0) \\
 &\equiv \tilde{R}(t_k, t_{k-1}) \dots \tilde{R}(t_2, t_1)\tilde{R}(t_1, t_0)R(t_0).
 \end{aligned}$$

Thus, multiplying by  $\begin{pmatrix} I & 0 \\ 0 & 0 \end{pmatrix}$ , one obtains backward error statements and subsequent forward error estimates for the case of the reduced QR factorization, obtaining bounds which cannot be worse than the bounds obtained on the full  $Q$ .

- (3) We remark that in this work we made use of the assumption of integral separation for the linear system. This is reasonable, since this is the generic case. However, it would be interesting to obtain direct bounds also removing this assumption, in a similar way to [14] Theorem 4.3.
- (4) For the general case of integration of a matrix equation of the form

$$\dot{y} = f(t, y), \quad y(t_0) = y_0 \text{ orthogonal}$$

we note that for the solution to remain orthogonal  $y^T \dot{y}$  must be skew-symmetric, hence  $f(t, y) = y s(t, y)$  where  $s(t, y)$  is skew-symmetric. Obtaining a result similar in flavor to Lemma 4.1 would require determining a splitting

$$\dot{y} = yL + y[s(t, y) - L], \quad y(t_0) = I,$$

where  $L$  is such that  $y[s(t, y) - L]$  remains small enough for  $y(t) \approx I$ . In the specific case considered here the splitting was motivated by integral separation, a natural property for linear nonautonomous differential equations  $\dot{x} = A(t)x$ , and we took advantage of the fact that the skew-symmetric matrix function  $S(Q, A)$  is linear in  $A$ .

**APPENDIX.** Here we derive some technical expressions that are useful in proving Lemma 4.1. In particular, we derive expressions relatively to  $q_{ij}^D(t, \hat{Q}, \omega)$ ,  $q_{ij}^T(t, \hat{Q}, \omega)$ , and  $q_{ij}^F(t, \hat{Q}, \omega)$  in (4.8) that are useful in obtaining the bounds (4.9, 4.10, 4.11) in the proof of Lemma 4.1.

First, consider  $q_{ij}^D(t, \hat{Q}, \omega)$ . We have, writing  $\hat{Q}(t) = [\hat{Q}_1(t) | \cdots | \hat{Q}_n(t)]$ ,

$$\begin{aligned} q_{ij}^D(t, \hat{Q}, \omega) - q_{ij}^D(t, I, \omega) &= q_{ij}^D(t, \hat{Q}, \omega) = \hat{Q}_{ij}[D_{ii} - D_{jj}] + [-\sum_{k=1}^{j-1} \hat{Q}_{ik} \hat{Q}_k^T + \sum_{k=j+1}^n \hat{Q}_{ik} \hat{Q}_k^T] D \hat{Q}_j \\ &= \hat{Q}_{ij}[D_{ii} - D_{jj}] + \sum_{l=1}^n D_{il} [\hat{Q}_{lj} \cdot \{-\sum_{k=1}^{j-1} \hat{Q}_{ik} \hat{Q}_{lk} + \sum_{k=j+1}^n \hat{Q}_{ik} \hat{Q}_{lk}\}] \\ (A-1) \quad &= D_{ii} [\hat{Q}_{ij} \cdot \{1 - \sum_{k=1}^{j-1} \hat{Q}_{ik}^2 + \sum_{k=j+1}^n \hat{Q}_{ik}^2\}] \\ &+ D_{jj} [-\hat{Q}_{ij} + \hat{Q}_{jj} \cdot \{-\sum_{k=1}^{j-1} \hat{Q}_{ik} \hat{Q}_{jk} + \sum_{k=j+1}^n \hat{Q}_{ik} \hat{Q}_{jk}\}] \\ &+ \sum_{l \neq i, j} D_{il} [\hat{Q}_{lj} \cdot \{-\sum_{k=1}^{j-1} \hat{Q}_{ik} \hat{Q}_{lk} + \sum_{k=j+1}^n \hat{Q}_{ik} \hat{Q}_{lk}\}]. \end{aligned}$$

By orthogonality we have

$$(A-2) \quad 1 = \sum_{k=1}^n \hat{Q}_{ik}^2 \quad \text{and} \quad \hat{Q}_{ij} \hat{Q}_{jj} = -\sum_{k \neq j} \hat{Q}_{ik} \hat{Q}_{jk},$$

so for  $i < j$ ,

$$\begin{aligned} q_{ij}^D(t, \hat{Q}, \omega) &= D_{ii} [\hat{Q}_{ij} (\hat{Q}_{ij}^2 + 2 \sum_{k=j+1}^n \hat{Q}_{ik}^2)] \\ (A-3) \quad &+ D_{jj} [-\hat{Q}_{ij} (1 - \hat{Q}_{jj}^2) + 2 \hat{Q}_{jj} \sum_{k=j+1}^n \hat{Q}_{ik} \hat{Q}_{jk}] \\ &+ \sum_{l \neq i, j} D_{il} [\hat{Q}_{lj} (-\sum_{k=1}^{j-1} \hat{Q}_{ik} \hat{Q}_{lk} + \sum_{k=j+1}^n \hat{Q}_{ik} \hat{Q}_{lk})]. \end{aligned}$$

Next, consider the term  $q_{ij}^T(t, \hat{Q}, \omega)$ . We have for  $i < j$  (and similarly for  $i > j$ )

$$\begin{aligned}
 (A-4) \quad q_{ij}^T(t, \hat{Q}, \omega) &= q_{ij}^T(t, I, \omega) = q_{ij}^T(t, \hat{Q}, \omega) = \sum_{k=1}^n \hat{Q}_{ik}(S(\hat{Q}, T))_{kj} \\
 &= -\sum_{k=1}^{j-1} \hat{Q}_{ik}(\hat{Q}_j^T T \hat{Q}_k) + \sum_{k=j+1}^n \hat{Q}_{ik}(\hat{Q}_k^T T \hat{Q}_j) \\
 &= -\sum_{k=1}^{j-1} \hat{Q}_{ik} \sum_{l=1}^n \sum_{m=l+1}^n \hat{Q}_{jl} T_{lm} \hat{Q}_{km} + \sum_{k=j+1}^n \hat{Q}_{ik} \sum_{l=1}^n \sum_{m=l+1}^n \hat{Q}_{kl} T_{lm} \hat{Q}_{jm} \\
 &= -\hat{Q}_{ii} \sum_{l=1}^n \sum_{m=l+1}^n \hat{Q}_{jl} T_{lm} \hat{Q}_{im} \\
 &\quad - \sum_{k=1, k \neq i}^{j-1} \hat{Q}_{ik} \sum_{l=1}^n \sum_{m=l+1}^n \hat{Q}_{jl} T_{lm} \hat{Q}_{km} + \sum_{k=j+1}^n \hat{Q}_{ik} \sum_{l=1}^n \sum_{m=l+1}^n \hat{Q}_{kl} T_{lm} \hat{Q}_{jm} \\
 &= -\hat{Q}_{ii} [\hat{Q}_{jj} \sum_{m=j+1}^n T_{jm} \hat{Q}_{im} + \hat{Q}_{ii} \sum_{l=1}^{i-1} \hat{Q}_{jl} T_{li} + \sum_{l=1, l \neq j}^n \sum_{m=l+1, m \neq i}^n \hat{Q}_{jl} T_{lm} \hat{Q}_{im}] \\
 &\quad - \sum_{k=1, k \neq i}^{j-1} \hat{Q}_{ik} \sum_{l=1}^n \sum_{m=l+1}^n \hat{Q}_{jl} T_{lm} \hat{Q}_{km} + \sum_{k=j+1}^n \hat{Q}_{ik} \sum_{l=1}^n \sum_{m=l+1}^n \hat{Q}_{kl} T_{lm} \hat{Q}_{jm}
 \end{aligned}$$

Finally, consider the term  $q_{ij}^F(t, \hat{Q}, \omega)$ . Using (1.5), we have for  $i < j$  (and similarly for  $i > j$ )

$$\begin{aligned}
 (A-5) \quad q_{ij}^F(t, \hat{Q}, \omega) - q_{ij}^F(t, I, \omega) &= [\hat{Q}S(\hat{Q}, F) - S(I, F)]_{ij} \\
 &= -\sum_{k=1}^{j-1} \hat{Q}_{ik} S(\hat{Q}, F)_{jk} + \sum_{k=j+1}^n \hat{Q}_{ik} S(\hat{Q}, F)_{kj} + F_{ji} \\
 &= F_{ji}(1 - \hat{Q}_{jj} + \sum_{k \neq i, k=1}^n \hat{Q}_{ik}^2 \hat{Q}_{jj}) \\
 &\quad - \hat{Q}_{ii} \sum_{(l,m) \neq (j,i), l,m=1}^n \hat{Q}_{lj} F_{lm} \hat{Q}_{mi} + \sum_{k=1, k \neq i}^n \hat{Q}_{ik} S(\hat{Q}, F)_{kj} .
 \end{aligned}$$

## REFERENCES

- [1] L. Ya. Adrianova, *Introduction to Linear Systems of Differential Equations*, Translations of Mathematical Monographs Vol. 146, AMS, Providence, R.I. (1995).
- [2] M. P. Calvo, A. Iserles, and A. Zanna. Numerical solution of isospectral flows. *Math. Comp.*, 66:1461–1486, 1997.
- [3] E. Celledoni and B. Owren. A class of intrinsic schemes for orthogonal integration. *SIAM J. Numer. Anal.*, 40:2069–2084, 2002.
- [4] X.-W. Chang, C. C. Paige and G. W. Stewart, Perturbation analyses for the  $QR$  factorization, *SIAM J. Matrix Anal. Appl.* **18** (1997), no. 3, 775–791.
- [5] X.-W. Chang and C. C. Paige, Componentwise perturbation analyses for the  $QR$  factorization, *Numer. Math.* **88** (2001), no. 2, 319–345.
- [6] A. Davey. An automatic orthonormalization method for solving stiff BVPs. *J. Comput. Phys.*, 51:343–356, 1983.
- [7] F. Diele, L. Lopez, and R. Peluso. The Cayley transform in the numerical solution of unitary differential systems. *Advances in computational mathematics*, 8:317–334, 1998.
- [8] L. Dieci, R. D. Russell, and E. S. Van Vleck. Unitary integrators and applications to continuous orthonormalization techniques. *SIAM J. Numer. Anal.*, 31:261–281, 1994.
- [9] L. Dieci, R. D. Russell, and E. S. Van Vleck. “On the computation of Lyapunov exponents for continuous dynamical systems”, *SIAM J. Numer. Anal.*, 34:402–423, 1997.
- [10] L. Dieci and E. Van Vleck. Computation of orthonormal factors for fundamental solution matrices. *Numerische Mathematik*, 83:599–620, 1999.
- [11] L. Dieci and E.S. Van Vleck, “Lyapunov Spectral Intervals: Theory and Computation,” (2002) *SIAM J. Numer. Anal.* **40** pp. 516–542.
- [12] L. Dieci and E.S. Van Vleck, “Lyapunov and Sacker-Sell Spectral Intervals,” (2007) *J. Dyn. Diff. Eq.* **19** pp. 265–293.
- [13] L. Dieci and E.S. Van Vleck, “On the Error in Computing Lyapunov Exponents by QR Methods,” (2005) *Numer. Math.* **101** pp. 619–642.

- [14] L. Dieci and E.S. Van Vleck, "Perturbation theory for approximation of Lyapunov exponents by QR methods," (2006) *J. Dyn. Diff. Eq.* **18** pp. 815–840.
- [15] J. K. Hale, *Ordinary Differential Equations*, (1980) Krieger.
- [16] D. Higham. Time-stepping and preserving orthonormality. *BIT*, 37:24–36, 1997.
- [17] W. Liu and E.S. Van Vleck, "Exponential Dichotomy for Asymptotically Hyperbolic Two-Dimensional Linear Systems," (2007) *submitted*.
- [18] G. H. Meyer. Continuous orthonormalization for boundary value problems. *J. Comp. Phys.*, 62:248–262, 1986.
- [19] K.J. Palmer, "The structurally stable systems on the half-line are those with exponential dichotomy," *J. Diff. Eqn.* **33** (1979), pp. 16–25.
- [20] K. J. Palmer, "Exponential dichotomy, integral separation and diagonalizability of linear systems of ordinary differential equations," *J. Diff. Eqn.* **43** (1982), pp. 184–203.
- [21] K. J. Palmer, "Exponential separation, exponential dichotomy and spectral theory for linear systems of ordinary differential equations," *J. Diff. Eqn.* **46** (1982), pp. 324–345.
- [22] G. W. Stewart, Perturbation bounds for the *QR* factorization of a matrix, *SIAM J. Numer. Anal.* **14** (1977), no. 3, 509–518.