SIAM J. NUMER. ANAL.
Vol. 46, No. 3, pp. 1166–1189

ON THE ERROR IN QR INTEGRATION*

LUCA DIECI† AND ERIK S. VAN VLECK‡

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 finding 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 classification. 65L

DOI. 10.1137/06067818X

Notation. An $(n \times n)$ real matrix X is indicated by $X \in \mathbb{R}^{n\times n}$. $\mathrm{diag}(X)$ is the matrix comprising the diagonal part of X, the rest being all 0's; $\mathrm{upp}(X)$ is the matrix comprising the upper triangular part of X, the rest being all 0's; and $\mathrm{low}(X)$ is the matrix comprising 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

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

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. For example, 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, 12, 18]. Our own interest in these techniques originates with methods to approximate Lyapunov exponents of

*Received by the editors December 20, 2006; accepted for publication (in revised form) November 19, 2007; published electronically March 7, 2008. This work was supported in part under NSF grants DMS-0139895, DMS-0139824, and DMS-0513438.
http://www.siam.org/journals/sinum/46-3/67818.html
†School of Mathematics, Georgia Institute of Technology, Atlanta, GA 30332 (dieci@math.gatech.edu).
‡Department of Mathematics, University of Kansas, Lawrence, KS 66045 (evanvleck@math.ku.edu).

dynamical systems, a feat which is greatly simplified when the system is brought in upper triangular structure; e.g., see [13, 8, 9].

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, i.e., 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 a linear time varying system, as well as global errors on other spectral quantities.

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

An outline 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 [10]. 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 in 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 comprises only 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 $X_0 = Q_0R_0$ 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:

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

where we have set

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

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

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

where

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

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 $\widetilde{B}$ we have $\widetilde{B}_{ij} = \left(Q^TAQ\right)_{ij} + \left(Q^TAQ\right)_{ji}$ for $i < j$ and $\widetilde{B}_{ii} = \left(Q^TAQ\right)_{ii}$, that is, $\widetilde{B} = \mathrm{upp}(Q^TAQ) + (\mathrm{low}(Q^TAQ))^T$.

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 includes [2, 3, 14, 12, 7, 16]. We are not going to review these works in detail, 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 $X(t_k) = Q(t_k)R(t_k)$, $k = 0, 1, 2, \ldots$. In other words, we are looking for the change of variables, the factor Q in the QR-factorization of X, at the grid-points $0 = t_0 < t_1 < \cdots$. Practically, the grid-points $\{t_k\}$ 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

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

where

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

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

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

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 $X(t_k)$,

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

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 the product of local triangular transition matrices:

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

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

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

where $\widetilde{B}$ 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 them Q_k's, are orthogonal. The key fact, which we proved in [10], 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

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

and at the same time

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

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 Theorems 3.1 and 3.16."

We will henceforth simply write

$$\|E_j\| \ \le\ \eta,\ \ j = 1, 2, \dots, \tag{2.7}$$

and stress that η 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 [10, 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*

$$\dot{\widehat{R}} = \big(B(t) + F(t)\big)\widehat{R},\ \ t \ge 0, \tag{2.8}$$

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

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

and F is bounded as

$$\sup_{t\ge 0}\|F\| \le c\eta + O(\eta^2) =: \delta. \tag{2.10}$$

For the sake of completeness, we remark that in (2.10) the bounds on the norm of F are obtained locally, on each subinterval $[t_{j-1}, t_j]$, $j = 1, 2, \ldots$, so that one really has $\sup_{t_{j-1}\le t\le t_j}\|F\| \le c_j\eta + O(\eta^2)$, 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

$$\sup_{t_{j-1}\le t\le t_j}\|F\| \;\approx\; \mathtt{TOL}(1+\kappa_{j-1}h_{j-1}),\;\; h_{j-1} = t_j - t_{j-1}, \tag{2.11}$$

where TOL is the local error for the obtained approximation to the transition matrix, and κ_{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 [10], we needed to have a certain condition satisfied; see [10, Assumption 3.5]. This amounted to the requirement that

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

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

3. Global error bounds for Q. Next, consider the unperturbed and perturbed triangular systems

$$\dot R = B(t)R,\; R(0) = R_0 \qquad \text{and} \qquad \dot{\widehat R} = [B(t)+F(t)]\widehat R,\; \widehat R(0) = R_0, \tag{3.1}$$

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

In [11], we proved (see 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 $R_D = \mathrm{diag}(R)$ and $R_U = \mathrm{upp}(R)$, so that $R = (I + R_U R_D^{-1})R_D =: ZR_D$. Accordingly, we have the unperturbed and perturbed diagonal systems

$$\dot R_D = D(t)R_D,\;\; \dot{\widehat R}_D = [D(t)+E(t)]\widehat R_D, \tag{3.2}$$

where $D(t) = \mathrm{diag}(B(t))$, $E = Z^{-1}FZ$, and $\widehat R = Z\widehat R_D$. Define

$$\mathrm{cond}(Z) = \sup_{t\ge 0}\|Z(t)\|\cdot\|Z^{-1}(t)\|, \tag{3.3}$$

and assume that $\mathrm{cond}(Z)$ 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 [8, Theorem 5.1].

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

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

for all $t, s,\ t \geq s \geq 0$, and $i = 1, 2, \ldots, n-1$. We also recall that integral separation is a generic property for linear systems (see [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

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

We make note here that the integral separation constants used in Lemmas 3.1 and 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

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

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,

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

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

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

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

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

where $a = \widetilde{a}$, $d \leq \widetilde{d} + 4M_{i,i+1}h_{max}$, $M_{i,i+1} = \sup_{t\geq 0} |B_{ii}(t) - B_{i+1,i+1}(t)|$, and $h_{max} = \sup_j (t_{j+1} - t_j)$. In other words, the problem with B also 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 ω is small enough. We proved this result in [11] under the assumption of integral separation of both unperturbed and perturbed triangular systems.

LEMMA 3.1 (see [11]). *Let* $\sup_{t\geq 0}\|B(t)\| = M$, *and let* a *and* d *be as defined in* (3.4). *Let* $\widehat{Q}$ *be the orthogonal factor in the QR-factorization of* $\widehat{R}_D$.

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

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

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

As an immediate consequence we have the following.

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

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

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

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

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. $\mathrm{cond}(Z)\rho < 1/2$, *and*
3. $\|\widehat{Q}(t_k) - I\| \leq \rho$.

Then we have

$$\|Q_k - Q(t_k)\| \leq \epsilon := \frac{3\mathrm{cond}(Z)\rho}{1 - 2\mathrm{cond}(Z)\rho},\ k = 0, 1, 2\ldots. \tag{3.12}$$

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

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

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

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

Thus, (3.10) may be written as

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

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

$$\|Q_k - Q(t_k)\| = \|Q^T(t_k)Q_k - I\| = \|W\| \le \epsilon. \qquad \square \tag{3.16}$$

It is possible to improve the perturbation bounds on the QR-factors of nearby matrices; see, e.g., [5, 4] 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 $\mathrm{cond}(Z)$. 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

$$\dot{\widehat{R}} = [B(t) + F(t)]\widehat{R}, \;\; t \ge 0, \;\; \widehat{R}(0) = R_0, \tag{4.1}$$

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

Below, we show that there exists an orthogonal change of variables to the upper triangular structure, that is, a change of variables $\widehat{R} = \widehat{Q}U$ with $\widehat{Q}$ orthogonal and U upper triangular with positive diagonal, such that $\widehat{Q}$ remains, under reasonable conditions, a small perturbation of the identity given the initial condition $\widehat{Q}(0) = \widehat{Q}_0 = I$. The proof of the lemma below uses a similar technique to that used to prove Lemma 3.1 (see [11]), 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 $\widehat{Q}$, we assume bounds on the entries of B and assume integral separation constants for both consecutive and nonconsecutive diagonal diagonal elements of B. Our bounds will be of the type $|\widehat{Q}_{ij}(t)| \le \rho_{ij}$, $i \ne j$, with $\rho_{ij} = \alpha_{ij}K_{ij}\omega$; see below. The key to obtaining this result is that the α_{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* $D = \mathrm{diag}(B)$ *and* $T = \mathrm{upp}(B)$. *Also, let* $\sup_{t\ge 0} \|F(t)\| \le \omega$. *Then there exists an orthogonal change of variables* $\widehat{Q}$, *with* $\widehat{Q}(0) = I$, *which bring* $B + F$ *to upper triangular structure* $C := \widehat{Q}^T[B + F]\widehat{Q} - \widehat{Q}^T\dot{\widehat{Q}}$.

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

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

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

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

and let $\alpha_{ij} = \alpha_{ji}$, $i > j$.

Set

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

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* $|\widehat{Q}_{ij}(t)| \le \rho_{ij}$ *for* $i \ne j$ *and all* $t \ge 0$, *where* $\rho_{ij} = \alpha_{ij}K_{ij}\cdot\omega$.

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

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

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

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

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

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

$$\begin{aligned}|q_{ij}(t,\widehat{Q},\omega)| &\le |q_{ij}(t,\widehat{Q},\omega) - q_{ij}(t,I,\omega)| + |q_{ij}(t,I,\omega)|\\ &\le \eta(\{\rho_{kl}\},\omega)\rho_{ij} + N(\omega),\end{aligned} \tag{4.7}$$

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

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

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

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

$$\begin{aligned}&|q_{ij}^D(t,\widehat{Q},\omega) - q_{ij}^D(t,I,\omega)|\\ &\quad\le \kappa_{ii}\left[\rho_{ij}\left(\rho_{ij}^2 + 2\sum_{k=j+1}^n \rho_{ik}^2\right)\right] + \kappa_{jj}\left[\rho_{ij}\sum_{k=1,k\ne j}^n \rho_{jk}^2 + 2\sum_{k=j+1}^n \rho_{ik}\rho_{jk}\right]\\ &\qquad + \sum_{l\ne i,j}\kappa_{ll}\left[\rho_{lj}\left(\sum_{k=1}^{j-1}\rho_{ik}\rho_{lk} + \sum_{k=j+1}^n \rho_{ik}\rho_{lk}\right)\right]\end{aligned}$$

$$
\begin{aligned}
&\le \kappa_{ii}\left[\rho_{ij}\left(\rho_{ij}^2+2\sum_{k=j+1}^{n}\rho_{ik}^2\right)\right]+\kappa_{jj}\left[\rho_{ij}\sum_{k=1,k\neq j}^{n}\rho_{jk}^2+2\sum_{k=j+1}^{n}\rho_{ik}\rho_{jk}\right]\\
&\quad+\sum_{l\neq i,j}\kappa_{ll}\left[\rho_{lj}\left(\rho_{li}+\rho_{il}+\sum_{k=1,k\neq i,j,l}^{n}\rho_{ik}\rho_{lk}\right)\right]\\
&=\rho_{ij}\left\{\omega^2\left[\kappa_{ii}\left(\beta_{ij}^2+2\sum_{k=j+1}^{n}\beta_{ik}^2\right)+\kappa_{jj}\sum_{k=1,k\neq j}^{n}\beta_{jk}^2+\sum_{l\neq i,j}\kappa_{ll}\frac{\beta_{lj}}{\beta_{ij}}\sum_{k=1,k\neq i,j,l}^{n}\beta_{ik}\beta_{lk}\right]\right.\\
&\quad\left.+\omega\left[\frac{2\kappa_{jj}}{\beta_{ij}}\sum_{k=j+1}^{n}\beta_{ik}\beta_{jk}+\sum_{l\neq i,j}\kappa_{ll}\frac{\beta_{lj}}{\beta_{ij}}(\beta_{li}+\beta_{il})\right]\right\}\\
&=:\rho_{ij}\eta_{ij}^{D}=:\rho_{ij}\big(\omega^2\eta_{ij}^{D,2}+\omega\eta_{ij}^{D,1}\big).
\end{aligned}
\tag{4.9}
$$

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

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

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

we obtain from (A.5)

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

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

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

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

and finally from (4.7) we obtain

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

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

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

or

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

or equivalently

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

Then we need to have

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

which we rewrite as

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

where in particular

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

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 ω satisfying $f(\omega) < 0$. This is guaranteed by (4.3).

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

What one expects from Lemma 4.1 are $\rho_{ij} \equiv \alpha_{ij} K_{ij} \omega$, where ρ_{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 $t \geq s$

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

as opposed to

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

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

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

and

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

The following theorem shows 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 $\widetilde{B}$ be the upper triangular matrix function obtained for the exact Q: $\widetilde{B} = Q^T AQ - S(Q,A)$. If we write the piecewise constant upper triangular B as $B = \widetilde{B} + \widetilde{F}$, then the upper triangular C that results from having only approximated Q satisfies for $\overline{Q} = Q\widehat{Q}$*

$$\begin{aligned} C &= \overline{Q}^T[A+G]\overline{Q} - S(\overline{Q}, A+G) \\ &= \widetilde{B} + \mathrm{upp}[\overline{Q}^T(A+G)\overline{Q} - Q^TAQ] + (\mathrm{low}[\overline{Q}^T(A+G)\overline{Q} - Q^TAQ])^T \\ &= B + \mathrm{upp}[-\widetilde{F} + \overline{Q}^T(A+G)\overline{Q} - Q^TAQ] + (\mathrm{low}[-\widetilde{F} + \overline{Q}^T(A+G)\overline{Q} - Q^TAQ])^T, \end{aligned} \tag{4.21}$$

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

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

Proof. We have

$$\begin{aligned} C &= \widehat{Q}^T[B+F]\widehat{Q} - S(\widehat{Q}, B+F) = \widehat{Q}^T[\widetilde{B} + \widetilde{F} + F]\widehat{Q} - S(\widehat{Q}, \widetilde{B} + \widetilde{F} + F) \\ &= \overline{Q}^T[A+G]\overline{Q} - S(\overline{Q}, A+G) = \mathrm{upp}[\overline{Q}^T(A+G)\overline{Q}] + (\mathrm{low}[\overline{Q}^T(A+G)\overline{Q}])^T \\ &= \widetilde{B} + \mathrm{upp}[\overline{Q}^T(A+G)\overline{Q} - Q^TAQ] + (\mathrm{low}[\overline{Q}^T(A+G)\overline{Q} - Q^TAQ])^T \\ &= B - \widetilde{F} + \mathrm{upp}[\overline{Q}^T(A+G)\overline{Q} - Q^TAQ] + (\mathrm{low}[\overline{Q}^T(A+G)\overline{Q} - Q^TAQ])^T \\ &= B + \mathrm{upp}[-\widetilde{F} + \overline{Q}^T(A+G)\overline{Q} - Q^TAQ] + (\mathrm{low}[\overline{Q}^T(A+G)\overline{Q} - Q^TAQ])^T \\ &= B + \mathrm{upp}[-\widetilde{F} + \overline{Q}^T(A+G)\overline{Q} - Q^TAQ] + (\mathrm{low}[-\widetilde{F} + \overline{Q}^T(A+G)\overline{Q} - Q^TAQ])^T, \end{aligned} \tag{4.22}$$

where we have used that $\mathrm{upp}[\widetilde{F}] = \widetilde{F}$ and $\mathrm{low}[\widetilde{F}] = 0$. 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 $\mathrm{upp}[S(Q,A)] + (\mathrm{low}[S(Q,A)])^T = 0$, 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 [11]) together with Theorem 3.2 of [11] 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, 8, 9]) the $\widetilde{B}$ problem is integrally separated, and hence so is the B problem. Besides the bound on F, δ, and the measure of integral separation, the K_{ij}, the bounds obtained depend on the $\kappa_{ij} = \sup_t |B_{ij}(t)|$ for $i \le j$.

(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, δ, 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 κ_{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 ρ_{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 ω_+. 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 ∞-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* $(n-1)(\widetilde{\rho} + \widetilde{\rho}^2) \le 1$. *Then,* $\|\widehat{Q}(t) - I\| \le \rho \equiv (n-1)(\widetilde{\rho} + \widetilde{\rho}^2)$ *and* $\|\widehat{Q}(t) - I\|_F \le \rho_F \equiv \sqrt{2(n^2-n)}\widetilde{\rho}$ *for all* $t \ge 0$. *Moreover, for* $k = 0, 1, 2, \ldots$, $\|Q_k - Q(t_k)\| \le \rho$.

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 ρ_{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,* $K_{n-1} \le K_{n-2} \le \cdots \le K_1$, *and* $\sum_{k=1}^{n-1} \rho_k + \rho_k^2 \le 1$, *then* $\rho_j \le \alpha_j K_j \omega$, $\|\widehat{Q}(t) - I\|_F \le (2\sum_{k=1}^{n-1}(n-k)\rho_k^2 + \sum_{i=1}^{n}(\sum_{k=1}^{i-1}\rho_k^2 + \sum_{i=1}^{n-i}\rho_k^2)^2)^{1/2}$, *and*

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

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

$$\alpha_j > 1 + (n-j-1)\kappa K_{j+1}\big(1 + (n-j-2)\kappa K_{j+2}(1 + \cdots (1 + \kappa K_{n-1}\alpha_{n-1}))\cdots\big) \tag{4.24}$$

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

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

$$\begin{aligned} &\alpha_{n-1} > 1, \ \ \alpha_{n-2} > 1 + \kappa K_{n-1}\alpha_{n-1} > \alpha_{n-1}, \\ &\alpha_{n-3} > 1 + 2\kappa K_{n-2}\alpha_{n-2} > \alpha_{n-2}, \ldots \end{aligned} \tag{4.25}$$

or in general

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

The result then follows by recalling that $\rho_j \le \alpha_j K_j \omega$ and obtaining the bounds on the matrix norms in a straightforward fashion using

$$\|\widehat{Q}(t) - I\| \le \sqrt{\|\widehat{Q}(t) - I\|_1 \|\widehat{Q}(t) - I\|_\infty}. \qquad \square$$

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 ω_+, as well as ρ, are different in these two contexts; notice the use of $\bar{\rho}$ in Corollary 3.1 versus the use of $\widetilde{\rho}$ in Corollary 4.1.

To estimate ω_+, 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 ω_+ in (4.4) suggest that

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

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 κ 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 κ 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 ω_+ 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 ρ_{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 $\mathrm{cond}(Z)$ when using the diagonalizing transformation. Indeed, it is interesting to compare the bounds one obtains with the two different approaches: (i) using the diagonalizing transformation Z, and (ii) working directly with the triangular system. We do this 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 $B = \left(\begin{smallmatrix} D_{11} & T_{12} \\ 0 & D_{22} \end{smallmatrix}\right)$.

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 ω in the two cases.

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

Then we have a bound $|\sin(\theta(t))| \le \rho := 2K\omega$, where (i) $|\sin(\theta(t))| = |\widehat{Q}_{12}(t)|$ and $\omega = \delta \mathrm{cond}(Z)$ 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

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

and

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

respectively.

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

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

and

$$\begin{aligned} \frac{1}{4K^2\kappa_{12}}\left[1-\frac{M}{K\kappa_{12}^2}\right] &= \frac{1-M/(K\kappa_{12}^2)}{4K^2\kappa_{12}} < \omega_+^T < \frac{M/(K\kappa_{12}^2)}{4MK/\kappa_{12}} \\ &= \frac{1}{4K^2\kappa_{12}} \le \frac{1}{4K\sqrt{\mathrm{cond}(Z)-2}}. \end{aligned} \tag{4.31}$$

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 κ_{ij}'s that measure the nonnormality 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 [8] 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:

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

We have the following result.

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

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

Then, for $t \ge s$, we have

$$\int_s^t p(r)dr \ge a(t-s) - d. \tag{4.33}$$

Proof. That $d \ge 0$ follows from

$$c(H) \ge c(H/2) \ge c(H/4) \ge \cdots \ge c(H/2^N) \ge M.$$

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

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

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

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

(4.34)

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 $[0,T]$:

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

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

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

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

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

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

and

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

$$\times \sum_{k=0, a_k = \epsilon h_k^{-1}}^{N-1} e^{-\left[\sum_{l=k+1}^{N-1} X_l\right]} h_k e^{-h_k Y_k} =: K(T), \tag{4.39}$$

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 $s \in [t_k, t_{k+1}]$,

$$\int_s^{t_{k+1}} p(r)dr \geq (t_{k+1}-s)Y_k = (t_{k+1}-s)\left[\frac{\epsilon}{h_k} - \frac{(\epsilon - h_k Y_k)}{h_k}\right] \geq \frac{\epsilon}{h_k}(t_{k+1}-s) - (\epsilon - h_k Y_k). \tag{4.40}$$

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

$$\int_s^{t_{k+1}} p(r)dr \geq (t_{k+1} - s)Y_k = a_k(t_{k+1} - s). \tag{4.41}$$

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

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

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 effectiveness of the bounds on the error in the orthogonal matrix function Q, in the following example, also considered in [11]. We report only 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

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

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

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

The parameter κ changes the degree to which there is nonnormality in the upper triangular part, and the parameters λ and ζ 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

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

TABLE 5.1
Error in the approximate Q varying the degree of nonnormality and integral separation, method, and tolerance.

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

κ	ζ	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

where

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

and

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

Results for this problem were obtained using the code LESLIS, which we developed (see www.math.gatech.edu/∼dieci and www.math.ku.edu/∼evanvleck).

In particular, we employ the continuous QR method (Cont QR in Table 5.1) using the projected fifth 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 (Disc QR in Table 5.1) with a fifth 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 TOL = 1.E-6 throughout. Before reporting on the results, we remark that (see (2.11)) we expect to have ω 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 varying κ and ζ values. We report on the error in Q in the two norm (the largest singular value of the error) at grid-points. 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 $|\widehat{Q}_{ij}(t)|$, in particular the α_{ij}, are somewhat difficult to give in closed form, the recursion (4.3) is straightforward to code. Likewise, the formula for ω_+, 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 κ_{ij}, the measure of nonnormality. 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 Lemmas 4.3 and 4.4 when $\zeta \geq 5/2$.

TABLE 5.2
Bounds on K_{23} obtained using Lemmas 4.3 and 4.4.

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

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

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

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 ρ_1 and ρ_2 obtained by using these bounds and those for α_i, K_i, $i = 1, 2$, for different values of κ and ζ.

At this point, we make an important observation: *If we use $\omega =$ `TOL` in the expressions obtained for ρ_1 and ρ_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 ω. 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 ω_+ 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 ω_+ (see the discussion after (4.27)).

Now, by comparison of the true errors obtained in Table 5.1, using the estimates for ω_+ from Table 5.3, and still adopting the rationale that $\omega =$ `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 κ, 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.

TABLE 5.3
Bound on the error in Q is dominated by $\rho_1 = \alpha_1 K_1 \omega$.

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

κ	ζ	ρ_1/ω	ρ_2/ω	$(2\rho_1+\rho_2)/\omega$	$\omega_+ \approx [(2\kappa_D+\kappa)\alpha_1^2 K_1^2]^{-1}$
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
10^2	1	7.7E1	1.4E0	1.6E2	1.6E-6
10^2	2	2.5E2	1.5E0	5.0E2	1.6E-7
10^2	4	1.2E4	2.0E0	2.4E4	7.4E-11

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 nonautonomous 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, e.g., 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. However, dealing with this case is—in principle—simple, since as a consequence of the combined backward and forward error analyses for all k we have

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

Thus, multiplying by $\left(\begin{smallmatrix} I & 0 \\ 0 & 0 \end{smallmatrix}\right)$, 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 [11, Theorem 4.3].

(4) For the general case of integration of a matrix equation of the form

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

we note that in order 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

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

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 relative to $q_{ij}^D(t,\widehat{Q},\omega)$, $q_{ij}^T(t,\widehat{Q},\omega)$, and $q_{ij}^F(t,\widehat{Q},\omega)$ 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 $q_{ij}^D(t,\widehat{Q},\omega)$. We have, writing $\widehat{Q}(t) = [\widehat{Q}_1(t)|\cdots|\widehat{Q}_n(t)]$,

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

By orthogonality we have

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

so for $i < j$,

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

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

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

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

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

REFERENCES

[1] L. Ya. Adrianova, *Introduction to Linear Systems of Differential Equations*, Transl. Math. Monogr. 146, AMS, Providence, RI, 1995.

[2] M. P. Calvo, A. Iserles, and A. Zanna, *Numerical solution of isospectral flows*, Math. Comp., 66 (1997), pp. 1461–1486.

[3] E. Celledoni and B. Owren, *A class of intrinsic schemes for orthogonal integration*, SIAM J. Numer. Anal., 40 (2002), pp. 2069–2084.

[4] X.-W. Chang and C. C. Paige, *Componentwise perturbation analyses for the QR factorization*, Numer. Math., 88 (2001), pp. 319–345.

[5] X.-W. Chang, C. C. Paige, and G. W. Stewart, *Perturbation analyses for the QR factorization*, SIAM J. Matrix Anal. Appl., 18 (1997), pp. 775–791.

[6] A. Davey, *An automatic orthonormalization method for solving stiff BVPs*, J. Comput. Phys., 51 (1983), pp. 343–356.
[7] L. Dieci and E. Van Vleck, *Computation of orthonormal factors for fundamental solution matrices*, Numer. Math., 83 (1999), pp. 599–620.
[8] L. Dieci and E. S. Van Vleck, *Lyapunov spectral intervals: Theory and computation*, SIAM J. Numer. Anal., 40 (2002), pp. 516–542.
[9] L. Dieci and E. S. Van Vleck, *Lyapunov and Sacker-Sell spectral intervals*, J. Dynam. Differential Equations, 19 (2007) pp. 265–293.
[10] L. Dieci and E. S. Van Vleck, *On the error in computing lyapunov exponents by QR methods*, Numer. Math., 101 (2005), pp. 619–642.
[11] L. Dieci and E. S. Van Vleck, *Perturbation theory for approximation of Lyapunov exponents by QR methods*, J. Dynam. Differential Equations, 18 (2006), pp. 815–840.
[12] L. Dieci, R. D. Russell, and E. S. Van Vleck, *Unitary integrators and applications to continuous orthonormalization techniques*, SIAM J. Numer. Anal., 31 (1994), pp. 261–281.
[13] 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 (1997), pp. 402–423.
[14] F. Diele, L. Lopez, and R. Peluso, *The Cayley transform in the numerical solution of unitary differential systems*, Adv. Comput. Math., 8 (1998), pp. 317–334.
[15] J. K. Hale, *Ordinary Differential Equations,* Robert E. Krieger, Huntington, NY, 1980.
[16] D. Higham, *Time-stepping and preserving orthonormality*, BIT, 37 (1997), pp. 24–36.
[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. Comput. Phys., 62 (1986), pp. 248–262.
[19] K. J. Palmer, *The structurally stable systems on the half-line are those with exponential dichotomy*, J. Differential Equations, 33 (1979), pp. 16–25.
[20] K. J. Palmer, *Exponential dichotomy, integral separation and diagonalizability of linear systems of ordinary differential equations*, J. Differential Equations, 43 (1982), pp. 184–203.
[21] K. J. Palmer, *Exponential separation, exponential dichotomy and spectral theory for linear systems of ordinary differential equations*, J. Differential Equations, 46 (1982), pp. 324–345.
[22] G. W. Stewart, *Perturbation bounds for the QR factorization of a matrix*, SIAM J. Numer. Anal., 14 (1977), pp. 509–518.